THE UNIVERSITY
of EDINBURGH

Using Numerical Simulations to Identify Observational Signatures of Self-Gravitating Protostellar Discs

Cassandra Hall, MPhys (HONS)

Doctor of Philosophy
The University of Edinburgh
February 2017

Abstract

In this thesis, I study numerical and semi-analytical models of self-gravitating protostellar discs, with the aim of furthering our understanding of the role of disc-self gravity in planet formation. At the time of writing, the ALMA era of observational astronomy is upon us. Therefore, I place my research into this context with synthetic images of both numerical and semi-analytical models.

I begin with an examination into the apparent lack of convergence, with increasing resolution, of the fragmentation boundary in Smoothed Particle Hydrodynamics (SPH) simulations of a protostellar disc. I run a suite of SPH with different numerical implementations, and find that even very similar implementations can fundamentally change the final answer.

I analyse a suite of SPH simulations that fragment to form gravitationally bound objects, with the motivation of informing future population synthesis model development. I find that fragment-fragment and fragment-disc interaction dominates the orbital evolution of the system even at very early times, and any attempt to produce a population of objects from the gravitational instability process must include these interactions.

Before a disc fragments, it will go through a self-gravitating phase. If the disc cools globally on a timescale such that it is balanced by heating due to gravitational stresses, the disc will be in a state of quasi-equilibrium. So long as the disc mass is sufficiently low, and spirals are sufficiently tightly wound, then angular momentum transport can be described by the local approximation, for which there is an analytical description.

Using this analytical description, I develop an existing 1D model into 3D, and examine a wide range of parameter space for which disc self-gravity produces significant non-axisymmetry. Using radiative transfer calculations coupled with synthetic observations, I determine that there is a very narrow range of parameter space in which a disc will have sufficiently large gravitational stresses so as to produce detectable spirals, but the stresses not be so large as to cause the disc to fragment. By developing a simple analytical prescription for dust, I show that this region of parameter space can be broadened considerably. However, it requires grains that are large enough to become trapped by pressure maxima in the disc, so I conclude that if self-gravitating spiral arms are detected in the continuum, it is likely that at least some grain growth has taken place.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified. Science is collaborative, as such, it is important to note the support and guidance of my collaborators, listed below, in obtaining the work shown in this thesis.

Prof. Ken Rice, Dr. Duncan Forgan, Prof. Tim Harries, Dr. Pamela Klaassen, Dr. Beth Biller, Dr. Ruobing Dong, Dr. Eugene Chiang.

Work presented in this thesis has been published, submitted for publication, or will be published, as follows:

1. Hall C., Forgan D., Rice K., Harries T. J., Klaassen P. D., Biller B., 2016, **Directly observing continuum emission from self-gravitating spiral waves**, *MNRAS*, 458, 306 (Chapter 6).

2. Hall C., Forgan D., Rice K., 2017, **Identifying and Analysing Protostellar Disc Fragments in Smoothed Particle Hydrodynamics Simulations**,*Submitted* (Chapter 4 and Chapter 5).

3. Dong R., Hall C., Rice K., Chiang E., 2015, **Spiral arms in gravitationally unstable protoplanetary disks as imaged in scattered light**, *ApJ*, 812, L3 (Chapter 6).

4. Hall C., Forgan D., Rice K., Harries T. J., **Directly observing continuum emission from self-gravitating spiral waves - II. The importance of particle trapping**,*in preparation* (Chapter 7).

Cassandra Hall
28th February 2016

For Natalie

Acknowledgements

This thesis is an accumulation of work that has come together since I began my postgraduate studies in September 2013. Since then, many, many people have played an important role in my life, as such, I am sure to fail to thank all of them.

I begin with thanking my primary supervisor, Professor Ken Rice. I am grateful to him for his patience, wisdom, and excellent supervision of my work. He has just the right balance of careful supervision and freedom for me to develop my own ideas, while maintaining the necessary care to detail. My secondary supervisor, Dr. Duncan Forgan, I must thank for his wisdom, willingness to share ideas, and various debugging skills. I cannot thank either of them enough.

I would also like to thank Professor Simon Goodwin, who supervised me during my masters year and summer project at The University of Sheffield. If he had not noticed that I liked to code, my thesis would have looked very different, if it was extant at all. For introducing me to the world of computational astrophysics, I thank him most sincerely.

Over the last three years, I have had the opportunity to work with several brilliant scientists, all of whom have taught me more than I thought my brain could hold. Pamela Klaassen has made me believe wizards are real, and they come in the form of people who have completed an ALMA postdoc. Ruobing Dong and Eugene Chiang have astounded me with their insight, enthusiasm, and diversity of knowledge. Without Tim Harries, I don't think I would have managed a single radiative transfer calculation. The TORUS code is excellent, and I hope to continue to use it for many years more. For broadening my knowledge on all things exoplanets, I must thank Beth Biller, Mariangela Bonavita and Eric Lopez.

The warmth and friendship that everyone has shown me at The Royal Observatory has made my time in Edinburgh some of the best years of my life. I am sure to leave people out, but my office-mates, Derek, Jvos, and David, I must thank for "letting" me open the windows - even in a snow storm. I thank Ryan for a lot of things, not least our #gainz gym sessions, subsequently immediately undone by trips to the pub, Ali, for lending me his meteorite, Rosie, for making me glad

I'm not a vegan, Sheggy for delicious Iranian treats that I didn't know I needed in my life, David N for sharing nightmare sushi stories, Vanessa and Patrick for amazing evenings filled with home cooking, friends and laughter.

I must thank Mike Wilson, for always being a sparkly ray of sunshine-filled happiness, Alex for making me laugh almost every day, Rachel for making me laugh by being mean to me, Rapha for trying to get me to cook, Clem and Ben for listening to me sporadically debate ideas with myself, David C for words of encouragement, Sam T for his endless debugging wisdom and Paula for our morning chats and looking after all of the PhD students at the Observatory. I should also thank Joey for introducing me to the world of doggos and puppers, despite its effect on my productivity, and definitely a thank you to Bruce for providing duct tape for my shoes on numerous occasions (don't ask). Vasiliy, Owen, Tom, Dan, Adam, Fran and Joe, thanks for putting up with my ~~hostile occupation of~~ visits to your office, and also thanks to the whoever rescued my beloved Flotta. I'm not naming names.

I would also like to thank everyone I played football with at the observatory. I have enjoyed countless soirées at *Chez Lopez* with Luke, Mariangela, Agnes, and Eric, and hope to enjoy many more. I must thank Rob and Linda for never failing to tell me something interesting, and introducing me to the razzle dazzle world of television recordings. If there is anyone I have forgotten, I hope my gratitude has been evident regardless.

I would also like to thank everyone at Rick Young's Blackbelt academy, for their support on and off the mat. To Rick Young, it has been a real privilege to learn from a world champion who is so grounded and humble, and clearly so passionate about what they teach. To all my sparring partners, Matthew, Joel, Liam, Alan, Ali, Carmen, Karen, Ceci, Fiona, Violet, Clio, Abbi and anyone I may have missed - Oss!

I owe a huge debt to my father, Stephen, for his unfailing love and support, and unwavering belief in my ability to succeed (frequently against substantial evidence to the contrary) throughout my 27 years of existence. At least in part, my curiosity about the world is owed to our reading of encyclopaedias together on the living room floor, trips to museums, and sharing of books. My sense of humour may also be his fault, although he will fervently deny it. I would also like to thank my brother, Connor, for sharing aforementioned sense of humour, and generally making me look well-behaved by comparison.

To my Urban Family, who have never questioned my ability to do anything, I must say I owe the most heartfelt thanks for your patience over the last three

years with my odd work-life balance, and for your support in this challenging endeavour. My childhood friend, Bower, I would like to thank for keeping my feet on the floor, and words of encouragement when I needed it most. I thank Grace, likewise, for her words of support, although in future I ask her to keep her tequila hangovers to herself. I thank Darren for our bro dates, and making me laugh until my sides hurt. To Stevie, Wayne, Chapman, Bob and Emma, for keeping in touch over all these years. I would also like to thank Erwin Albert Sagan Lovelace Hamilton Prince Needlefeet LCB Kenny-Hall III for his excellent "little craggy boulder" impressions, and patience during his forced "Lion King" re-enactments. I am glad, although admittedly, surprised, that I still have hands.

Finally, and most importantly, I thank my wife, Natalie. I thank her for her patience, support, kindness, courage, and unfailing ability to make me laugh. During the last three years, she has been a pillar of emotional and physical support, and has cooked, baked, cleaned, washed and generally done every practical task while I have had my nose buried in a book. Without her, I would be lost, and I owe her the deepest debt of gratitude possible. She shines more brightly in my life than any celestial body. I am lucky to have her.

Contents

List of Figures

List of Tables

You know that the beginning is the most important part of any work.....

Plato, The Republic

1

Introduction

Discs are found on every rung of the cosmic ladder, from galaxies to planets, and are the inevitable consequence of angular momentum conservation. At the largest scale, active galactic nuclei (AGN) are powered through the accretion of cold matter onto the central supermassive black hole. Quasars, a type of AGN, are some of the most luminous objects in the universe; the first quasar ever discovered, 3C 273, has an absolute magnitude of -26.7 (Greenstein & Schmidt, 1964), corresponding to a luminosity of $\sim 2.5 \times 10^{40}$ W. Since the Sun's luminosity is $\sim 3.8 \times 10^{26}$ W, the quasar 3C 273 is ~ 14 orders of magnitude more luminous than the Sun!

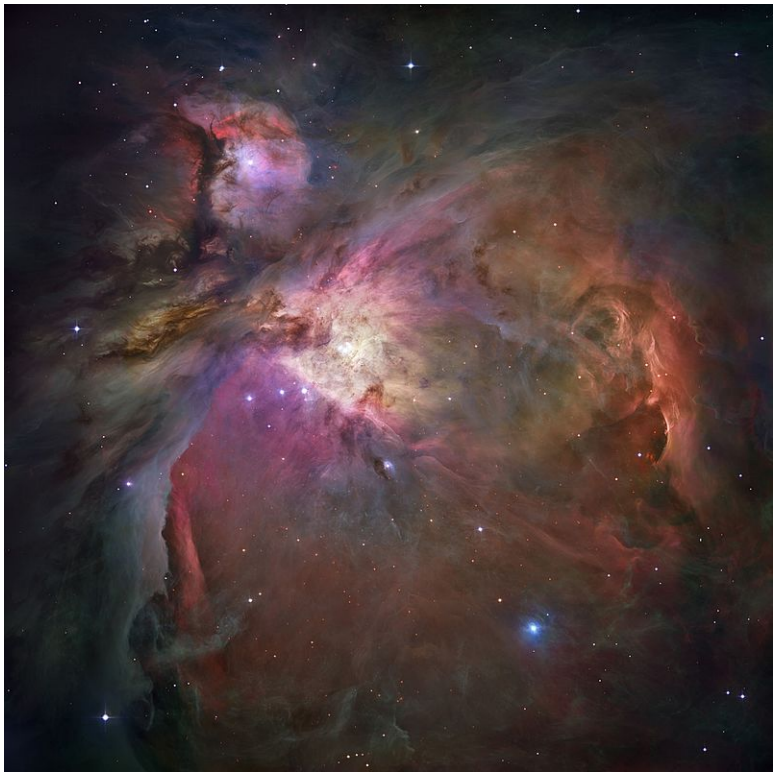


Figure 1.1: The Orion nebula, located within the Orion Molecular Cloud Complex, a region of massive star formation located ~ 400 pc from Earth. Image credit: NASA, ESA, M. Robberto (Space Telescope Science Institute/ESA) and the Hubble Space Telescope Orion Treasury Project.

At far smaller, much less energetic scales, we find *circumstellar discs*, which are more relevant to this thesis. These discs are found around young stars, which have not yet reached the hydrogen burning, or *main sequence* phase of their life. Since they surround a protostar, these discs are often called *protostellar discs*.

To understand the origin of a protostellar disc, we must consider star formation. Stars are born in turbulent, gas and dust rich regions of space called *giant molecular clouds* (GMCs), such as the Orion Nebula (shown in Figure 1.1), which is located in the Orion Molecular Cloud Complex.

Regions of the GMC, when sufficiently cool, will become gravitationally unstable and collapse to form stars. However, these regions are so large that even a small amount of velocity anisotropy generates significant angular momentum. As such, the collapse must proceed by maintaining rotation along some axis, which results in a central protostar, and surrounding protostellar disc.

These protostellar discs subsequently become the site of *planet formation*; birthplaces of extrasolar planetary systems that are, sometimes, wildly different from our own Solar System. The first ever detected extrasolar planetary system was around the pulsar[1] PSR1257+12 (Wolszczan & Frail, 1992). However, it is highly unlikely that these planets formed before the supernova. This is because the energy released in the supernova would unbind any planets already in orbit around the star. It is therefore most likely that these planets formed from the debris disc left *after* the supernova.

The discovery of the first extrasolar planet that could, potentially, have formed through the same mechanism as our own Solar System was planet 51 Peg b, a $\sim$ Jupiter mass planet orbiting a solar-type star, discovered by radial velocity monitoring (Mayor & Queloz, 1995). What was initially somewhat dismissed as an anomaly, and later proved to be astonishing, was that the orbit of 51 Peg b is smaller than the orbit of Mercury. With the discovery of similar planets, astronomers quickly realised that while these planets may not be common, they

[1]When a massive star undergoes a supernova at the end of its life, so long as the star is sufficiently massive (but not so massive as to form a black hole), a neutron star will be formed. This dense neutron star retains most of the angular momentum of the original formation material, so spins very rapidly. The neutron star emits beams of electromagnetic radiation. If these beams are misaligned with the spin axis of the star, it will appear to *pulse*.

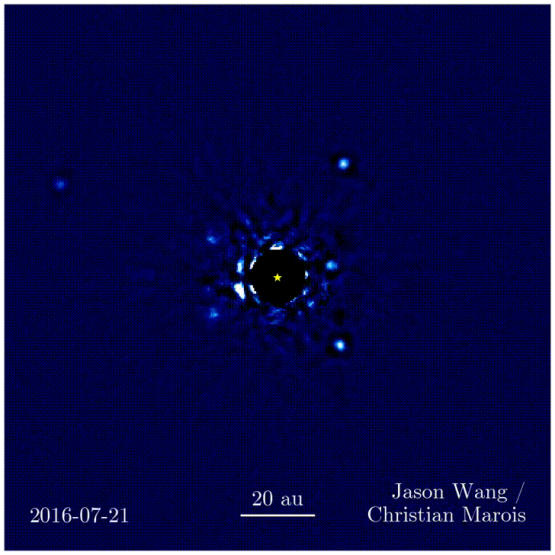


Figure 1.2: The directly imaged system HR 8799, showing four planets in the mass range $5 - 13$ M_J. This image was taken at the W. M. Keck Observatory in Hawaii. The planet closest to the central star has an orbital period of $\sim$ 40 years, and the furthest planet has an orbital period of $\sim$ 400 years. Image credit: Jason Wang (image), Christian Marois (data analysis), Quinn Konopacky (Orbit determination), Bruce Macintosh (data taking), Travis Barman (data taking), Ben Zuckerman (data taking).

are certainly not an anomaly.

This presented a significant challenge to planet formation theory, leading astronomers to investigate orbital migration and in-situ theories of formation (Batygin et al., 2016; Boley et al., 2016) that could explain these so-called *hot Jupiters*. Later, the discovery of the planetary system HR 8799 (Marois et al., 2008, 2010), shown in Figure 1.2, presented an almost opposite problem to the theory of planet formation: with projected separations of 14 au, 24 au, 38 au and 68 au, and masses estimated between 5 and 13 M_J, these observations were outside the semi-major range where astronomers thought planet formation would occur, according to the traditional *core accretion* theory of planet formation.

In the core accretion paradigm, growth begins at the microscopic level. Solid particles of dust collide, sticking together until they form planetesimals, which continue to grow through collisions until a protoplanetary core forms. This core is large enough to accrete a substantial gaseous envelope, ultimately forming a gas-giant planet. However, this model suffers from substantial *timescale* problems; that is, numerical core accretion models predict that giant planet formation should occur on timescales longer than the protoplanetary disc lifetime.

Dust particles feel no pressure force, and so orbit the central star at the Keplerian velocity. However, the gas is subject to an outwards pressure force, resulting in orbital speeds that are sub-Keplerian. This results in a drag on the solid dust particles, which lose angular momentum and spiral in towards the central star. These dust particles will either be destroyed by collisions with each other, since the relative velocities are sufficiently high that coagulation is unlikely, or destroyed as they approach the central star through photoevaporative processes. This inward migration is rapid, typically a few hundred years, which means that giant planet formation must be more rapid than outlined in the traditional core accretion paradigm.

One possibility is that these planets formed through direct gravitational collapse, known as *fragmentation*, due to a gravitational instability in the disc. At early times, these discs have a mass that is comparable with that of the central star. As such, the self-gravity of the disc plays an important role in its subsequent dynamical evolution. Fragmentation struggles to explain the existence of planets below a few M_J (see, e.g., Rice et al. 2015), simply because there is a fundamental lower limit, known as the *Jeans mass*, that a perturbation, whether in a disc or a

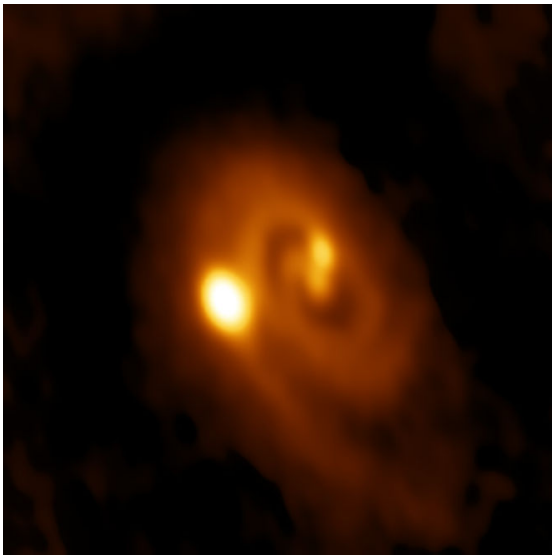


Figure 1.3: L1448 IRS 3B, the first disc ever to be caught in the act of fragmentation. It contains three protostars, the central pair are separated by 61 au, and the tertiary star is coincident with the spiral arm at a location of 183 au from the center of the disc. It is located in the Perseus molecular cloud, some $\sim$ 230 pc away. L1448 IRS 3B is classified as a Class 0 protostellar system, which means it is in the very earliest phase of the star formation process, deeply embedded in an envelope of accreting material. Image credit: B. Saxton (NRAO/AUI/NSF); ALMA ESO/NAOJ/NRAO.

molecular cloud, must reach before it is able to collapse. This is governed by the balance between the freefall timescale, t_{ff} and the sound speed, t_s of the region that is susceptible to collapse. As a region begins to collapse, the shock waves take a time t_s to cross the region, acting to restore the pressure balance in the system. Therefore, if $t_{ff} < t_s$, then the region will collapse. For this condition, we are able to define a density, and therefore a mass, that describes the minimum criteria for this to take place.

Despite difficulty in explaining the formation of planets with masses below a few M_J, fragmentation does, currently, remain a possible formation mechanism for planets such as those directly imaged around the HR 8799 system. While this

Figure 1.4: A long exposure image of a selection of ALMA antennae. The Milky Way runs through the center of the image, Mars is visible in the below centre left of the picture, and zodiacal light can be seen top right. Image credit: ESO/B. Tafreshi (twanight.org)

thesis was being written, a disc was captured, for the first time, in the act of fragmenting into gravitationally bound objects (Tobin et al., 2016), the triple protostar system L1448 IRS 3B, shown in Figure 1.3.

The system contains two central protostars, separated by a distance of 61 au, and a tertiary star coincident with a spiral arm in the outer disc, with a measured separation from the central disc of 183 au. Since the outer part of the disc has been determined to be susceptible to fragmentation, it was deemed that the tertiary protostar is likely to have formed through direct collapse in the outer part of the disc (Tobin et al., 2016). Although these objects are sufficiently massive as to be considered protostars rather than planets, it is, still, encouraging that this physical process has now been observed in Nature.

The image shown in Figure 1.3 was captured with the Atacama Large Millimeter/-submillimeter Array (ALMA) instrument, an array of telescopes in the Atacama desert of northern Chile. A selection of antennae, set against a backdrop of the Milky Way, is shown in Figure 1.4. Mars is visible in below centre left of the picture, and zodiacal light, sunlight scattered by dust in the zodiacal cloud, is visible in the top right corner. The majesty of the image is justified, since ALMA, with its unprecedented resolution and sensitivity, is ushering a new era of astronomical discovery into the annals of science. With this instrument, scientists hope that the secrets of star and planet formation will be revealed to humanity.

The focus of this thesis is gravitationally unstable protostellar discs. I use a mix of hydrodynamics simulations, semi-analytical models, radiative transfer calculations and synthetic imaging to characterise fragments formed through

gravitational instability of protostellar discs, and to outline observational signatures of gravitationally unstable discs.

In Chapter 2, I outline the important physics that governs the structure and evolution of protostellar discs. I discuss disc structure, evolution, and two complementary theories of planet formation. In Chapter 3, I outline the history and methodology of smoothed particle hydrodynamics (SPH). I then describe preliminary work that was carried out in the early stages of my PhD, investigating the lack of convergence of the fragmentation boundary with increasing resolution. At the time of conducting this work, this was a particularly important problem to solve, as a lack of convergence with increasing resolution may have implied that fragmentation could take place for very small perturbations, and, therefore, in any protostellar disc, given enough time. This would call our fundamental understanding of gravitational instability into question.

In Chapter 4, I introduce a new method for identifying fragments in SPH simulations, based on a spatial search of the density derivative (which I call DDS, which stands for density derivative search). I then compare this to an existing method, named `CLUMPFIND`, and discuss the relative merits of the two approaches, and how their strengths and weaknesses are complementary. In Chapter 5, I use the DDS to identify fragments in SPH simulations, and subsequently track them using a standard merger tree algorithm. I compare these fragments to gravitational instability population synthesis models, and suggest distributions for eccentricity and inclination that could be used to improve future population synthesis models. I examine the radial density and temperature profile of a selection of fragments, and discuss the accuracy of their representation in

population synthesis models as polytropic spheres.

In Chapter 6, I describe how I converted an existing 1D semi-analytic model of a quasi-steady, self-gravitating disc into a 3D model, which is then used as the geometry for radiative transfer calculations. From this, I generated multi-wavelength synthetic images of these discs as they would be observed using ALMA, and drew conclusions about the relatively narrow region of parameter space that high m-mode ($\gtrsim 8$) self-gravitating discs must exist in if they are to be observed with ALMA. This Chapter also includes work that was done in collaboration with Ruobing Dong, Ken Rice and Eugene Chiang. This is present in section 6.6, and I have been careful to distinguish between work that I produced, and work that is that of my collaborator.

Chapter 7 is the final science chapter of this thesis, and describes the role of dust enhancement in the detection of self-gravitating spiral waves. Building upon the semi-analytic model described in Chapter 6, I develop a prescription for dust enhancement that does not require time integration of the equations of motion. Despite its simplicity, the results are encouraging, widening the parameter space where non-axisymmetric structure due to disc self-gravity is detectable with ALMA.

Finally, in Chapter 8, I summarise the main results of this thesis, and discuss prospects for future work.

There is no point in using the word "impossible" to describe something that has clearly happened.

Douglas Adams, Dirk Gently's Holistic Detective Agency

2

Protostellar Discs, Planet Formation, and the Physics That Governs It All

2.1 Introduction

All stars that form today do so from gas that exists in giant clouds, composed of molecular gas and dust. These molecular clouds are dense regions of the interstellar medium (ISM), with typical densities of $\sim$ 100 molecules per cubic centimetre. The clouds are cold, typically of order $\sim$ 10 K or so, and as such the hydrogen gas is in its molecular form, H_2. This molecule emits spectral lines at

wavelengths of ∼1-100 μm, and due to water vapour in the Earth's atmosphere, it cannot be easily detected from the Earth's surface. Consequently, we rely on molecular tracers, such as the $J = 1 - 0$ transition of ^{12}CO and ^{13}CO to infer the masses of these clouds. In addition to mass, these molecular tracers provide us with kinematic information about the cloud, such as rotation, or regions of collapsing gas.

The most dense regions of molecular clouds are known as molecular cloud cores, and these regions typically have number densities of $\sim 10^4$ cm^3, and are the actual sites of star formation in the galaxy. Their usual scale is ~ 0.1 pc across, and measurements indicate low velocity gradients, typically of the order 1 km s^{-1} pc^{-1}, indicating that rotation is not dynamically important at these early times.

However, there is still substantial angular momentum in such a core, due to its large radial extent. Since the angular momentum vector, $\mathbf{L}$, is given by

$$\mathbf{L} = m(\mathbf{r} \times \mathbf{v}), \tag{2.1}$$

where m is mass, $\mathbf{r}$ is radial separation and $\mathbf{v}$ is velocity, we can see that so long as either m and/or $\mathbf{r}$ is large, then $\mathbf{L}$ will also be large, even if $\mathbf{v}$ is small.

Assuming that a core contains at least enough mass to form our Solar System gives a total mass of $\sim 10^{33}$ g. Using our velocity gradient, 1 km s^{-1} pc^{-1}, on a diameter of 0.1 pc gives us a total angular momentum of $J_{\rm core} \sim 10^{54}$ g cm^2 s^{-1}. For comparison, the total angular momentum of the sun is $\sim 10^{49}$ g cm^2 s^{-1}, and the total angular momentum of the entire solar system is $\sim 10^{50}$ g cm^2 s^{-1}.

This discrepancy between the angular momentum in the cloud core, and the angular momentum in the final solar system it forms, is known as the *angular momentum problem* of star formation. The solution to this problem is thought to be a combination of processes; star formation is by its nature dynamic, and thus angular momentum may be lost in outflows and interactions with nearby stars. However, what motivates our discussion here is the idea that a protostellar disc is the vessel through which much of this angular momentum may be redistributed.

By appealing to conservation of angular momentum, and noting that the *specific* angular momentum (i.e., per unit mass), $j_{\rm K}$, of a disc in Keplerian orbit around a 1 $\rm M_{\odot}$ star is given by

$$j_{\rm K} = \sqrt{\rm G M_{\odot} R}, \tag{2.2}$$

then equating $j_{\rm K}$ to the *specific* angular momentum in our molecular cloud, $j_c = J_{\rm core}/M_{\rm cloud} \approx 10^{21}$ cm^2 s^{-1}, we expect our protostellar discs to have an outer radius $R \sim 500$ au, assuming no angular momentum loss. This is justified by observations showing that typical sizes of protostellar discs around young stars are $\sim 100-1000$ au (Benisty et al., 2015a; Grady et al., 2009). Although magnetic fields are not considered in this thesis, it is prudent to note that the efficiency of magnetic braking[2] determines how extended a protostellar disc will be.

In this Chapter, I first discuss the main physical principles that govern the formation and evolution of protostellar discs. Armed with this, I then discuss

[2]The basic idea of magnetic braking is that a fast rotating disc is surrounded by a slowly rotating massive envelope, and the two are magnetically linked. Ionised gas in the disc will push on magnetic field lines, losing energy. This energy loss causes the gas to decrease in velocity, or "brake", and will fall onto the star since it is no longer rotationally supported. This suppresses disc formation, which requires the outward spread of material.

the conditions for collapse in molecular clouds, the subsequent classification of the young stellar objects that this collapse forms, and the disc structure and ultimate angular momentum transport.

Finally, I discuss current theories of planet formation in the context of protostellar discs.

2.2 Governing Physics

In this section, I outline the three areas of physics that govern the formation and evolution of protostellar discs, and ultimately planet formation. Since the majority of matter that makes up stars and planets is gas, I consider first of all the equations of fluid dynamics. Since it is gravity that makes our molecular cloud cores collapse, I next discuss gravity. Finally, for something to cool enough to collapse, a process must exist by which energy can be radiated away, or else the collapse will be halted. For this reason, I discuss some basic equations of radiative transfer. Magnetic fields undoubtedly play an important role in the formation of protostellar discs (see, e.g. Li et al. 2011), however, we do not consider them in this thesis.

2.2.1 Hydrodynamics

In this section, we develop the equations of motion in fluid mechanics, beginning with Euler's assumption of a perfect fluid, and then relaxing that assumption to derive the Navier-Stokes equation for an imperfect fluid. We begin with three basic principles:

1. Mass is conserved. It is neither created, nor destroyed, simply rearranged in space.

2. Momentum is conserved. The rate of change of momentum of a portion of the fluid is equal to the force applied to it. This is only true in systems in which mass is conserved.

3. Energy is conserved. It is neither created, nor destroyed, it simply moves from one form to another. It may be transferred from one region of space to another, but if we consider an enclosed system, the total energy in this system is conserved.

2.2.1.1 Continuity Equation

We begin by defining a region in space, D, that is filled with a fluid. Let $\mathbf{x}$ be a point in D, defined as $\mathbf{x} = (x, y, z)$. Now imagine a particle in this fluid, with a trajectory such that it passes through $\mathbf{x}$. It then has a velocity $\mathbf{v}(\mathbf{x}, t)$, and if time is fixed, this $\mathbf{v}$ is the *spatial velocity field* of the fluid in D.

We now assume that the fluid mass density can be described as $\rho(\mathbf{x}, t)$, then the mass, m, in a subregion of the fluid, W (this subregion is fixed in time), at time t, is determined by

$$m(W, t) = \int_W \rho(\mathbf{x}, t) dV, \tag{2.3}$$

where dV is the volume element we are considering. Now, the rate of change of mass in subregion W is

$$\frac{d}{dt} m(W, t) = \frac{d}{dt} \int_W \rho(\mathbf{x}, t) dV = \int_W \frac{\partial \rho}{\partial t}(\mathbf{x}, t) dV. \tag{2.4}$$

Let us assume this rate of change of mass describes the *inward* flow of mass. Since point 1 in our list tells us that mass must be conserved, this increase in mass must be equal to the flow of mass across the surface of W, into the volume element. This is known as the *flux*, and is by definition $\mathbf{j} = \rho\mathbf{v}$, where $\mathbf{v}$ is our velocity field as described earlier. The rate that this mass is flowing through our surface ∂W is then a surface integral, such that

$$\int\int_{\partial W} S\mathbf{j} \cdot d\mathbf{S}. \tag{2.5}$$

Since we have our vector field $\mathbf{v}$, there exists a unit normal vector to this field $\hat{\mathbf{n}}$. We can then rewrite equation 2.5,

$$\int\int_{\partial W} S\mathbf{j} \cdot d\mathbf{S} = \int\int_{\partial W} S\mathbf{j} \cdot \hat{\mathbf{n}} dS. \tag{2.6}$$

Now, since we have defined our surface as ∂W, and we know that our flux $\mathbf{j} = \rho\mathbf{v}$, we are now in a position to write down the *integral form of the law of conservation of mass*,

$$\frac{d}{dt}\int_W \rho dV = -\int_{\partial W} \rho\mathbf{v} \cdot \hat{\mathbf{n}} dA, \tag{2.7}$$

where we have made use of assuming our surface can be characterised as a 2

dimensional surface. If we make use of the divergence theorem[3], this gives us

$$\int_{\partial W} \rho \mathbf{v} \cdot \hat{\mathbf{n}} dA = \int_{W} \nabla \cdot (\rho \mathbf{v}) dV. \tag{2.8}$$

This then gives us

$$\int_{W} \left[\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) \right] = 0, \tag{2.9}$$

and since this must hold everywhere, then this gives us the **continuity equation**,

$$\boxed{\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0.} \tag{2.10}$$

2.2.1.2 Momentum Equation

We now appeal to point 2 in our list, the conservation of momentum. Let us consider a volume element of fluid. On that fluid are two different types of forces: body-type and stress-type. Stress-type forces, such as pressure, are applied to the *boundary* of the fluid. Body-type forces, such as gravity, act on the *interior* of the solid. If we now take the surface of our fluid, $S = \partial W$, then there exists a pressure with a function $\mathrm{P}(\mathbf{x}, t)$, which acts upon the surface such that the force across S per unit area, is given by:

$$\mathbf{F}_{\mathrm{S}} \text{ per unit area} = P(\mathbf{x}, t)\hat{\mathbf{n}}, \tag{2.11}$$

[3]The divergence theorem can be expressed as follows. Consider a region in space, W, with a boundary ∂W. The volume integral of the divergence, $\nabla \cdot \mathbf{F}$, of $\mathbf{F}$, over W, and the surface integral of $\mathbf{F}$, over the boundary ∂W, are related by

$$\int_{W} (\nabla \cdot \mathbf{F})\, dW = \int_{\partial W} \mathbf{F} \cdot d\mathbf{a}.$$

where $\hat{\mathbf{n}}$ is the normal vector to the surface. This can then be integrated for the total force on the fluid from the stress on the surface S,

$$\mathbf{F}_{\mathrm{S}} = -\int_{\partial W} P\hat{\mathbf{n}} dA, \tag{2.12}$$

where the integral is negative because the normal vector points outwards from the surface. We can then use the divergence theorem (described in footnote 3), defining $\mathbf{l}$ as any vector in space, then from

$$\mathbf{l} \cdot \mathbf{F}_{\mathrm{S}} = -\int_{W} (\nabla p) \cdot \mathbf{l} dV, \tag{2.13}$$

we can see that

$$\mathbf{F}_{\mathrm{S}} = -\int_{W} \nabla P dV. \tag{2.14}$$

The force per unit volume on any piece of fluid is then ∇P + *some body force per unit volume*, $\mathbf{F}_b$ (where $\mathbf{F}_b$, for example, is an external force such as gravity). Appealing again to point 2 in our list, we know that the force $\mathbf{F}$ on a body of mass m is determined by $\mathbf{F} = m\mathbf{a}$, where $\mathbf{a}$ is the acceleration. We can express this as a force per unit volume, so $\mathbf{F} = \rho\mathbf{a}$, which we now write as

$$\mathbf{F} = \rho\frac{d\mathbf{v}}{dt} = -\nabla P, \tag{2.15}$$

since we know from equation 2.14 that force per unit volume is $-\nabla P$. We are now in a position to note that since momentum is conserved, we know that the force felt by the material per unit volume, $\rho\mathbf{a}$, is equal to the force exerted on the

fluid by stress on its surface, plus any external body forces, so

$$\rho \mathbf{a} = -\nabla P + \mathbf{F}_b \tag{2.16}$$

We now make a distinction between a Lagrangian and Eulerian derivative. In Eulerian mechanics, we fix the location in space, and do not refer to a particular fluid element. However, a Lagrangian treatment follows the path of an individual particle in the fluid. Therefore, the particle velocity is now a function of time, as is the particle's position. In this case, the velocity is now

$$\mathbf{v} = \mathbf{v}(x(t), y(t), z(t), t) = (\dot{x}(t), \dot{y}(t), \dot{z}(t)). \tag{2.17}$$

We differentiate this with respect for time, and use the chain rule, to express the acceleration as

$$\mathbf{a}(t) = \frac{\partial \mathbf{v}}{\partial x}\dot{x} + \frac{\partial \mathbf{v}}{\partial y}\dot{y} + \frac{\partial \mathbf{v}}{\partial z}\dot{z} + \frac{\partial \mathbf{v}}{\partial t}. \tag{2.18}$$

Now, applying the definition of a Cartesian derivative,

$$\mathbf{v} \cdot \nabla \equiv \mathbf{v}_x \frac{\partial}{\partial x} + \mathbf{v}_y \frac{\partial}{\partial y} + \mathbf{v}_z \frac{\partial}{\partial z}, \tag{2.19}$$

to equation 2.18, we rewrite this as

$$\mathbf{a}(t) = \frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v}. \tag{2.20}$$

We note that this makes use of the *material derivative*

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla, \tag{2.21}$$

which takes into account the changing nature of the fluid over time, and is an important function in fluid mechanics. Now that we have our new expression for the acceleration of the fluid that takes into account this dynamic nature, we combine equations 2.16 and 2.20 to define the **momentum equation**

$$\boxed{\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}(\nabla P + \mathbf{F}_{\mathrm{b}}).} \tag{2.22}$$

2.2.1.3 Energy Equation

Having developed the **continuity equation**, based on the conservation of mass, and the **momentum equation**, based on the conservation of momentum, we are now in a position to derive the **energy equation**, based on our third point in our list, the conservation of energy. We begin by assuming that the total energy of the fluid can be written as

$$E_{\mathrm{total}} = E_{\mathrm{kinetic}} + E_{\mathrm{internal}}, \tag{2.23}$$

noting that in the case of an incompressible flow, all the energy is kinetic, since the rate of change of energy in a portion of incompressible fluid equals the rate at which pressure and body forces do work. We consider an isentropic fluid, where at least some of the work done by pressure and body forces alters the internal energy of the particle. A fluid is isentropic when its flow is both adiabatic (no heat is added to the flow), and reversible, so no energy is lost to friction or dissipation.

Defining u as *internal energy per unit mass*, then our expression for total energy

in a region of fluid W becomes

$$E_{\text{total}} = E_{\text{kinetic}} + E_{\text{internal}} \tag{2.24}$$

$$= \frac{1}{2}\int_W \rho|\mathbf{v}|^2\, dV + \int_W \rho u\, dV, \tag{2.25}$$

which we can simplify by allowing e to mean total energy per unit mass, and then we have $E = \rho e$, or

$$E = \frac{1}{2}\rho|\mathbf{v}|^2 + \rho u. \tag{2.26}$$

The rate of change of this energy is then

$$\frac{dE}{dt} = \frac{1}{2}\rho\frac{D|\mathbf{v}|^2}{Dt} + \rho\frac{du}{dt}, \tag{2.27}$$

where we can again use the *material derivative* defined in equation 2.21, and omitting the resulting algebra on $|\mathbf{v}|^2$ for brevity, we now have

$$\frac{dE}{dt} = \rho\left[\mathbf{v}\cdot\left(\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\right)\right] + \rho\frac{du}{dt}. \tag{2.28}$$

Now, beginning with the First Law of Thermodynamics, we have

$$dU = TdS - PdV, \tag{2.29}$$

where U is internal energy, T is temperature, S is entropy and V is volume. In the specific form, then we have

$$du = Tds - Pd\left(\frac{1}{\rho}\right) = TdS + \frac{P}{\rho^2}d\rho. \tag{2.30}$$

Since we are dealing with an isentropic flow, by its nature it is adiabatic, which means $dS = 0$. Taking the derivative with respect to t gives

$$\frac{du}{dt} = -\frac{P}{\rho^2}\frac{d\rho}{dt}, \tag{2.31}$$

and we can now use the continuity equation to derive the **energy equation** for an isentropic fluid,

$$\boxed{\frac{dE}{dt} = \rho\left[\mathbf{v}\cdot\left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v}\right)\right] - P(\nabla\cdot\mathbf{v})} \tag{2.32}$$

2.2.1.4 Navier Stokes Equation

Thus far, we have assumed that forces across a surface act only normal to that surface. In our definition of an ideal fluid, that is certainly the case. However, in a more general fluid, this is not true. In fact, it can be seen that the stress forces acting on a fluid are best described by a tensor known as the *stress tensor*, σ, which is shown in Figure 2.1. We can see that the stress has 9 components, and is thus a rank-2 tensor. The resultant stress vector, $\mathbf{T}$, in this case, has components in the three unit vectors depicted by $\mathbf{e}$. Since pressure is a special case of stress force, whereby the force acts normal to the surface, then we can rewrite our momentum equation in equation 2.22, (making use of the material derivative, for ease of understanding of physical motivation) as

$$\rho\frac{D\mathbf{v}}{Dt} = \nabla\cdot\sigma + \mathbf{F}_{\mathrm{b}}, \tag{2.33}$$

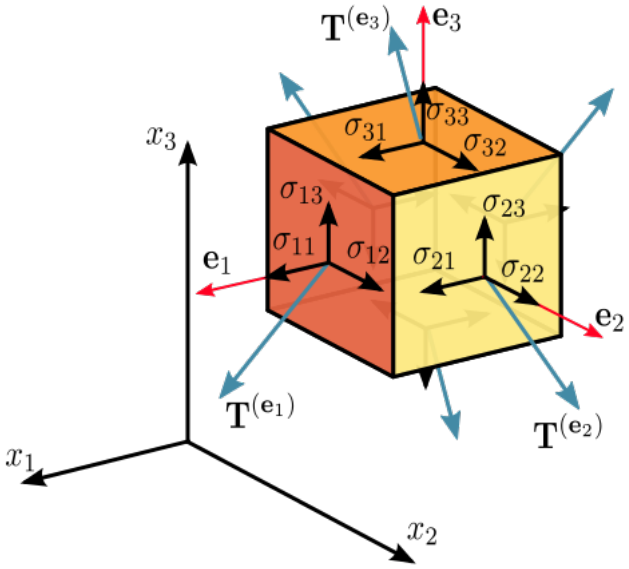


Figure 2.1: Depiction of the stress tensor, σ acting on an element of fluid in three dimensions. The resulting stress vector, $\mathbf{T}$, has components given by unit vectors $\mathbf{e}_1$, $\mathbf{e}_2$ and $\mathbf{e}_3$. Image credit: Wikipedia user Sanpaz, image distributed under a CC BY-SA 3.0 license.

where $\mathbf{F}_\mathrm{b}$ is a force acting on the body of the fluid, as before, such as gravity. Now, by examination of Figure 2.1, we can see that σ is composed of normal and shear components. Since the pressure is of particular interest to us, it makes sense to decompose our stress tensor into its pressure components and deviatoric

components (i.e., the components responsible for the distortion of the fluid), thus:

$$\begin{aligned}\sigma_{ij} &= \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{pmatrix} \\ &= -\begin{pmatrix} P_{11} & 0 & 0 \\ 0 & P_{22} & 0 \\ 0 & 0 & P_{33} \end{pmatrix} + \begin{pmatrix} \tau_{11}+P_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22}+P_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33}+P_{33} \end{pmatrix} \\ &= -P\mathbb{I} + \mathbb{T} \end{aligned} \tag{2.34}$$

Where $\mathbb{I}$ is the identity matrix (where the main diagonal are all ones, and zero elsewhere), and $\mathbb{T}$ is the deviatoric stress tensor. We can now write the **Navier-Stokes Equation** in its most general form,

$$\rho\frac{D\mathbf{v}}{Dt} = -\nabla P + \nabla\cdot\mathbb{T} + \mathbf{F}_b. \tag{2.35}$$

All of our uncertainty about the properties of the fluid is now contained in $\mathbb{T}$. To use our equation, we must make some assumptions about its property. For our purposes, we consider an incompressible, isotropic[4], viscous fluid that is considered Newtonian (i.e.the shear stress is linearly proportional to the shear strain, or the strain rate), in the fluid. If this is the case, then we have

$$\tau = \nu\frac{d\mathbf{v}}{dt}, \tag{2.36}$$

[4]Not to be confused with isentropic, which means a reversible process in which no heat is transferred, isotropic means a substance is uniform in all directions.

where ν is the coefficient of viscosity, giving for our stress tensor

$$\mathbb{T}_{ij} = \nu\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) + \delta_{ij}\lambda\nabla\cdot\mathbf{v}, \tag{2.37}$$

where λ is the second coefficient of viscosity. In our incompressible case, $\nabla\cdot\mathbf{v} = 0$, and so we do not need to worry about λ. In the interest of completeness, λ is related to the bulk viscosity of the material, and is generally set such that $\lambda = \frac{2}{3}\nu$, so that the coefficient of bulk viscosity is zero. However, this appears to only have some basis in theory of monatomic gases, and thus is somewhat controversial, even today.

Substituting our expression for our stress tensor in equation 2.37 back into equation 2.35, and using our complete expression for the material derivative, we finally arrive at the **Navier-Stokes Equation for an incompressible, Newtonian, viscous fluid**

$$\boxed{\frac{\partial\mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \nu\nabla^2\mathbf{v} + \mathbf{F}_b.} \tag{2.38}$$

2.2.2 Gravity

For an astronomer, gravity is the most important of the four fundamental forces of Nature. Arguably, gravity is the fundamental force that has allowed intelligent life to form, since I could not ponder my existence without a cloud of gas collapsing under its own gravity to form the Sun, and subsequently the Earth, some ~ 4.5 billion years ago. As of the time of writing, a *Theory of Everything*, or grand unified theory, that can link gravity with the other three fundamental forces

of nature eludes mankind. As it stands, quantum field theory explains the electromagnetic force, and the strong and weak nuclear force, but is incompatible with general relativity, our most well-accepted gravitational theory. For our purposes, we do not need to give gravity its full relativistic treatment. It is sufficient to simply assume that Newton's laws are true. Newton's theory of gravitation describes the attractive force, $\mathbf{F}$, between body 1 and body 2, such that

$$\mathbf{F} = \frac{\mathrm{G}m_1 m_2}{r^2}, \tag{2.39}$$

where G is the gravitational constant, m_1 and m_2 are the masses of body 1 and body 2, and r is the distance between them. A scalar expression relating force and gravitational potential energy U is simply

$$F = -\frac{dU}{dr}. \tag{2.40}$$

We can see from this that we can define the gravitational force field $\mathbf{g}$, such that

$$\mathbf{g} = -\nabla\Phi, \tag{2.41}$$

where ϕ is the gravitational potential field, and we can now update our Navier-Stokes equation to

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P + \nu\nabla^2\mathbf{v} - \nabla\Phi \tag{2.42}$$

2.2.2.1 Hydrostatic Equilibrium

If we consider a spherical cloud of mass, that mass will be in hydrostatic equilibrium when its internal pressure is supporting it against gravitational collapse. We begin by considering a spherical shell of mass dm, given by

$$dm = \rho \cdot \text{volume} = \rho \cdot 4\pi r^2 \, dr, \tag{2.43}$$

where ρ is the density of the shell, r is the distance from the centre of the sphere to the shell, and dr is the thickness of the shell. The gravitational force on this shell, F_{g}, is given by

$$\begin{aligned} F_{\text{g}} &= -\frac{\text{G}M(r)\, dm}{r^2} \\ &= -\frac{\text{G}M(r)\rho 4\pi r^2 \, dr}{r^2}, \end{aligned} \tag{2.44}$$

where G is the gravitational constant. The shell also experiences two pressure forces, one from the mass above it, pushing down, and one from the mass below it, pushing up. This change in pressure across the shell is given by

$$dP = P(r + dr) - P(r), \tag{2.45}$$

and since the force due to pressure is $F_{\text{P}} = \text{pressure} \cdot \text{area}$, then we have

$$F_{\text{P}} = dP \, 4\pi r^2. \tag{2.46}$$

For the cloud to remain stationary, these forces must balance, such that

$$dP\,4\pi r^2 = -\frac{\mathrm{G}M(r)\rho 4\pi r^2\,dr}{r^2}, \tag{2.47}$$

cancelling the $4\pi r^2$ terms, and rearranging, we have the condition for hydrostatic equilibrium in a spherically symmetric cloud:

$$\frac{dP}{dr} = -\frac{\mathrm{G}M(r)\rho}{r^2}. \tag{2.48}$$

2.2.2.2 The Jeans Instability: Gravitational Collapse

Mass collapses under its own gravitational attraction if the outwards pressure of the mass is no longer sufficient enough to support it. This only occurs if the *Jeans Criterion* is met. If this is the case, then the mass has exceeded the Jeans Mass, or the size has exceeded the Jeans length. We begin from the virial theorem, which states that the condition for equilibrium in a gravitationally bound, stable system is

$$2K + U = 0, \tag{2.49}$$

where K is kinetic energy, and U is the gravitational potential energy. The system will collapse if $2K < |U|$. Assuming the system under consideration is a uniform density sphere, then the gravitational potential energy is

$$U = \frac{3}{5}\frac{\mathrm{G}M^2}{R}. \tag{2.50}$$

The kinetic energy of our sphere, we can get from considering the kinetic energy of an individual particle,

$$K_i = \frac{3}{2} k_B T, \tag{2.51}$$

where T is temperature, and k_B is the Boltzmann constant. Therefore, for our sphere, the total kinetic energy is

$$K = \frac{3}{2} N k_B T, \tag{2.52}$$

where N is the number of molecules in the sphere. Since

$$N = \frac{M}{\mu m_{\mathrm{H}}}, \tag{2.53}$$

then we can write our condition for collapse as

$$\frac{3 M k_B T}{\mu m_{\mathrm{H}}} < \frac{3}{5} \frac{\mathrm{G} M^2}{R}. \tag{2.54}$$

Since the radius, R, of a uniform sphere of constant density, ρ, and mass, M, is given by

$$R = \left(\frac{3M}{4\pi\rho} \right)^{\frac{1}{3}}, \tag{2.55}$$

we can substitute this into our condition in equation 2.54, to obtain an expression for the *Jeans mass*

$$M_{\mathrm{J}} = \left(\frac{5 k_{\mathrm{B}} T}{\mathrm{G} \mu m_{\mathrm{H}}} \right)^{\frac{3}{2}} \left(\frac{3}{4\pi\rho} \right)^{\frac{1}{2}}, \tag{2.56}$$

and by using equation 2.55, the corresponding *Jeans Length*,

$$R_{\mathrm{J}} = \left(\frac{5\mathrm{k_B}T}{\mathrm{G}\mu m_{\mathrm{H}}} \cdot \frac{3}{4\pi\rho} \right)^{\frac{1}{2}}. \tag{2.57}$$

As this cloud collapses, gravitational energy is released. If all of that gravitational energy was converted to thermal energy, we can see by equation 2.56 that the Jeans mass will increase, halting collapse in that region. However, in the early stages of collapse, the cloud is optically thin to infrared radiation, and thus most of the photons escape the cloud, and the cloud essentially collapses in freefall, on a timescale given by

$$t_{\mathrm{ff}} = \sqrt{\frac{3\pi}{32\mathrm{G}\rho}}. \tag{2.58}$$

During the freefall collapse phase, the density increases, decreasing the Jeans mass, meaning that due to initial cloud inhomogeneities, subregions of the cloud undergo local collapse, or *fragmentation*. If there was no process by which this fragmentation would stop, regions and sub regions of those regions would continue to collapse. In reality, the increasing density in the collapsing regions eventually renders them optically thick to infrared photons, causing the cloud to heat. The collapse is therefore no longer isothermal, and proceeds adiabatically.

This adiabatic process causes the Jeans mass to *increase* with increasing density, essentially halting fragmentation, and providing a minimum mass threshold for star formation. This can be seen as follows. We write the adiabatic relationship between pressure and density

$$P = K\rho^{\gamma}, \tag{2.59}$$

where K is a constant and γ is the ratio of specific heats. The ideal gas law relates pressure, density and temperature thus:

$$P = \frac{\rho k_{\mathrm{B}} T}{\mu m_{\mathrm{H}}}, \tag{2.60}$$

which means we can express an adiabatic relationship between density and temperature such that

$$T = K_2 \rho^{\gamma - 1}, \tag{2.61}$$

where $K_2 = {}^{K\mu m_{\mathrm{H}}}/_{k_{\mathrm{B}}}$. Substituting our adiabatic relationship between temperature and density into our expression for Jeans mass in equation 2.56, and assuming $\gamma = {}^{5}/_{3}$, for atomic hydrogen, we can see that

$$M_{\mathrm{J}} \propto \rho^{\frac{1}{2}}, \tag{2.62}$$

which effectively acts to halt fragmentation. Now, we develop a criteria for minimum fragment mass by discussing a concept known as the *opacity limit*. We have already shown, in our initial definition of the Jeans mass, that in the isothermal regime,

$$M_{\mathrm{J}} \propto \sqrt{\frac{T^3}{\rho}}. \tag{2.63}$$

This leads us to the idea of a critical mass at the boundary of these two regimes. This can be estimated by making some simple considerations. First, the gravitational binding energy, B, of a collapsing ball of gas (assuming,

simplistically, constant density), is given by

$$B = -\frac{3}{5}\frac{\mathrm{G}M^2}{R}. \tag{2.64}$$

We know that this collapse happens on the freefall timescale, t_{ff}. Therefore, the power available per second to heat the gas is given by

$$\frac{|B|}{t_{\mathrm{ff}}} = \frac{3}{5\pi}\left(\frac{\mathrm{G}^3M^5}{R^5}\right)^{\frac{1}{2}}. \tag{2.65}$$

At the same time, the gas is cooling through the emission of blackbody radiation, at luminosity $L = e4\pi R^2\sigma T^4$, where e is the efficiency factor. To maintain isothermal collapse, we must have

$$\frac{|B|}{t_{\mathrm{ff}}} \lesssim L, \tag{2.66}$$

or

$$\frac{3}{5\pi}\left(\frac{\mathrm{G}^3M^5}{R^5}\right)^{\frac{1}{2}} \lesssim e4\pi R^2\sigma T^4. \tag{2.67}$$

Rearranging, this gives us an expression for critical mass,

$$M \lesssim M_{\mathrm{crit}} = \left(\frac{400}{9}\frac{\pi^4 e^2\sigma^2 T^8 R^9}{\mathrm{G}^3}\right)^{\frac{1}{5}} \tag{2.68}$$

which, if exceeded, will result in the conversion to adiabatic collapse, and subsequently halt the collapse. However, for the collapse to proceed, the mass

must be larger than the Jeans mass, so we have

$$M_{\mathrm{J}} \lesssim M \lesssim M_{\mathrm{crit}}. \tag{2.69}$$

If M_{J} and M_{crit} are equal, this is known as the *opacity limit*, and it describes the minimum fragment size that may form. Combining our expression for critical mass, with our expression for Jeans mass and Jeans length, and substituting in the values of constants, gives

$$M_{\mathrm{fragment}} \approx 1.13 \times 10^{31} \frac{T^{\frac{1}{4}}}{e^{\frac{1}{2}}}. \tag{2.70}$$

For a cloud at $T \sim 10$ K, and assuming that the cloud is irradiating as a blackbody at 100% efficiency, gives us a fragment mass of ~ 10 $\mathrm{M_J}$.

This approach is useful, and intuitive, when considering the collapse of astrophysical structures which are, at least approximately, spherical. However, the *Jeans instability*, or, more strictly, the *Toomre instability*, which is the generalisation of the Jeans instability to rotating structures, also governs the collapse of regions inside protostellar discs, and we discuss this later in this chapter.

2.2.3 Radiative Transfer

Our understanding of the universe, in large parts, comes from our assessment of radiation fields in some way or another. Electromagnetic radiation is a spectrum of components, with wavelength λ and frequency ν related by $c = \lambda\nu$, where c is the speed of light in a vacuum. Light propagates as a wave, but since the formulation of wave-particle duality, and the subsequent development of quantum

mechanics, it is known to also be a particle, called a photon. These photons propagate with an energy given by $E = h\nu$, where h is Planck's constant, 6.63×10^{-34} m^2 kg s^{-1}. In this section, we describe some quantities associated with radiative transfer, the equations of radiative transfer, and the numerical methods we have used in this thesis to solve these equations for our particular problem set.

2.2.3.1 Fundamental Quantities

Suppose an area dA is exposed to some radiation for a time dt. The amount of energy passing through that element is given by $dE = FdAdt$, where F is the flux, the amount of energy per unit time, per unit area. For a spherical isotropic source, this flux falls off in an inverse square manner, since energy must be conserved. If we consider two concentric, spherical shells at radius r_1 and r_2, where $r_2 > r_1$, enclosing a source, then the luminosity passing through them is given by

$$F_1 4\pi r_1^2 = F_2 4\pi r_2^2. \tag{2.71}$$

Rearranging this gives

$$F_2 = \frac{F_1 r_1^2}{r_2^2}, \tag{2.72}$$

then assuming the first shell remains constant, we have

$$F_2 \propto \frac{1}{r_2^2}. \tag{2.73}$$

This equation is enough to calculate the flux of stars, and other large bodies, and it measures the amount of flux over all frequencies. Telescopes are our eyes to

the universe, and the only way we can make sense of what we see is to decompose this information into spectral frequencies. How much of the radiation is in X-rays is particularly important if we are looking at high energy accretion discs around black holes. However, we are unlikely to see X-rays emitting from a protostellar disc at $\sim$ 50 K. It becomes helpful, therefore, to consider radiation in given frequency intervals. The energy, dE, crossing an area, dA, in a time dt, and in the frequency range $d\nu$ is given by

$$dE = I\, dA\, dt\, d\nu\, d\Omega, \tag{2.74}$$

where $d\Omega$ is solid angle, and I is what is referred to as the **specific intensity**, which is a quantity that is conserved along a ray path in empty space. For this reason, the infinitesimal frequency range $d\nu$ is specified in the definition. Similarly, dE is also a conserved quantity.

We can see that the specific intensity then has units

$$I = \frac{energy}{time \cdot area \cdot solid\, angle \cdot frequency}, \tag{2.75}$$

so will change with space, time, direction (θ and Ω) and frequency. This is why, without powerful computers, astronomers have been unable to solve radiative transfer problems that vary in three dimensions. This specific intensity is a function of seven variables, and we discuss numerical methods below in section 2.2.4.

If we now consider an area dA placed inside a radiation field, such that there is

radiation in all directions, then the **net flux** in a given frequency range is given by

$$F_\nu = \int I_\nu \cos\theta \, d\Omega, \tag{2.76}$$

where θ is the angle between the normal to the surface of the area and the direction of the ray. The $\cos\theta$ term reflects the decrease in effective area with increasing θ. We can now update our expression for energy in equation 2.74 to the more general case where the ray is not normal to the surface

$$dE = I \, \cos\theta \, dA \, dt \, d\nu \, d\Omega. \tag{2.77}$$

We are now in a position to discuss the equation of radiative transfer.

2.2.3.2 The Equation of Radiative Transfer

If we consider a cylinder of radiation travelling through a medium, three processes govern the changes in that cylinder.

1. **Emission**. The medium through which the cylinder of radiation is travelling will emit radiation itself, which will add to the total intensity in the cylinder.

2. **Absorption**. The medium through which the cylinder of light is travelling will absorb some of the radiation, causing the total intensity to decrease in the cylinder.

3. **Scattering**. Radiation is scattered in and out out of the cylinder, adding and subtracting to the intensity respectively. For simplicity, we do not

consider scattering here.

Emission We define the *emission coefficient* j as the energy emitted per unit time, per unit solid angle, per unit volume, in a frequency range $d\nu$ as

$$dE = j_\nu \, dV \, d\Omega \, dt \, d\nu. \tag{2.78}$$

Assuming the emission is isotropic, then

$$j_\nu = \frac{1}{4\pi} P_\nu, \tag{2.79}$$

where P_ν is power per unit volume, per unit frequency. If we now equate equation 2.74 and 2.78, we see

$$dI_\nu \, dA = j_\nu \, dV, \tag{2.80}$$

noting that j is assumed constant along the path of the cylinder. Since $dV = dAdS$, where dS is path length, then we have the intensity added to the beam by spontaneous emission,

$$dI_\nu = j_\nu \, dS. \tag{2.81}$$

Absorption Absorption causes a loss of intensity in the beam, and is quantified by the *absorption coefficient*, α_ν (with units of length^{-1}), given by

$$dI_\nu = -\alpha_\nu I_\nu \, dS. \tag{2.82}$$

This can be understood by rewriting

$$\alpha_\nu = n\sigma_\nu, \tag{2.83}$$

where n is the number of absorbers per unit volume, and σ_ν is the effective area of the absorber. In astronomy, this is most commonly written as

$$\alpha_\nu = \rho\kappa_\nu, \tag{2.84}$$

where ρ is the mass density, and κ_ν is the opacity coefficient (of units area mass^{-1}).

The Radiative Transfer Equation Along a ray, by combining equations 2.81 and 2.82, we can see that the variation of specific intensity is given by

$$\frac{dI_\nu}{dS} = -\alpha_\nu I_\nu + j_\nu. \tag{2.85}$$

This is known as *the equation of radiative transfer*. However, the emission coefficient j_ν is more difficult to work with, in practice, than its more helpful relative S_ν, the *source function*, defined as

$$S_\nu \equiv \frac{j_\nu}{\alpha_\nu}. \tag{2.86}$$

To see this, we introduce the concept of *optical depth*, defined by

$$d\tau_\nu = \alpha_\nu \, ds, \tag{2.87}$$

or in integral form

$$\tau_\nu = \int_{s_0}^{s} \alpha_\nu \, ds. \tag{2.88}$$

The integral, τ_ν, indicates if a photon is able to travel along a path through a medium without being absorbed. If $\tau > 1$, the medium is optically thick to the photon, and will typically be absorbed before it escapes. If the material is optically thin to the photon, then $\tau_\nu < 1$ and the photon will typically escape the medium without being absorbed.

Dividing equation 2.85 by α now gives us the equation of radiative transfer in the form

$$\frac{dI_\nu}{d\tau_\nu} = -I_\nu + S_\nu, \tag{2.89}$$

which is solved by the use of integrating factors to give

$$I_\nu(\tau_\nu) = I_\nu(0)e^{-\tau_\nu} + \int_0^{\tau_\nu} S(\tau_\nu)e^{-\tau_\nu} \, d\tau_\nu. \tag{2.90}$$

In general, this equation is complicated, requiring advanced numerical methods to solve it (see section 2.2.4). However, there are two simplifying cases that are worth mentioning. First, in the case of a blackbody emitter, then $I_\nu = B_\nu$, i.e. the intensity is simply given by the Planck function. However, this is only true if the body is optically thick to the radiation everywhere. If the body is thermally emitting, then $S_\nu = B_\nu$, but I_ν does not equal B_ν unless the matter is optically thick.

2.2.4 Numerical Methods in Radiative Transfer

If we wish to generate theoretical observations of astrophysical systems, such as molecular clouds or protostellar discs, their radiation fields must be accurately determined. To do this requires solving the equation of radiative transfer numerically, by solving derivatives, or integrals, by differences or summation respectively. Historically, this was only possible for one dimensional (1D), plane parallel cases, as this high level of symmetry meant techniques like matrix inversion could solve the resulting set of equations. Since we do not consider any 1D geometries in this thesis , we do not describe these 1D methods. However, for the interested reader, we refer them to the works of Wick (1943), which is published in German, and the further development by Chandrasekhar (1947).

Astronomy, in this century, has moved on from considering 1D plane parallel cases. In the era of space telescopes and interferometric arrays, we are availed of a wealth of images, measurements and spectra for highly non-symmetric structures, such as protostellar discs, or even protostellar discs that are in the process of fragmenting to become bound objects (Tobin et al., 2016). With this lack of geometrical simplicity, iterative methods such as the *Monte Carlo* (MC) approach become increasingly attractive. The beauty of MC methods for solving integrals is their simplicity, meaning that complex, multivariable problems are reducible to an iterative program format to solve. This is shown in Figure 2.2, which shows some arbitrary function which is not possible to integrate analytically. The *unknown* area under the curve is described by the integral

$$I = \int_0^b f(x)dx, \tag{2.91}$$

where 0 and b are the limits of the function. These limits describe an area of *known* size, shown by the rectangle in Figure 2.2, with that area, A, given by

$$A = f(x)_{\max}(x = b) \tag{2.92}$$

Armed only with a random number generator, equations 2.91 and 2.92, we proceed as follows to solve the integral:

1. Generate two random numbers, $0 < z_1 < 1$ and $0 < z_2 < 1$.
2. Get a random point inside A, by doing $x_z = z_1 b$, $f_z = z_2 f(x)_{\max}$.
3. Is this point under the curve? (i.e. $f_z < f(x)$)?
   - **Yes**: Accept: $N_{\text{accept}} = N_{\text{accept}} + 1$. (Ticks in Figure 2.2)
   - **No**: Reject! (Crosses in Figure 2.2).
4. Repeat steps 1-3 N times. Larger N will give better accuracy.

Once this is complete, our integral is then given by

$$I = \frac{N_{\text{accept}}}{N} A. \tag{2.93}$$

This method can be intuitively understood in terms of probability, if the random points are more likely to fall under the curve of the function, than outside the curve of the function, then the ratio I/A will be large. If the curve of the function is equal to our rectangle, then this probability is 1.

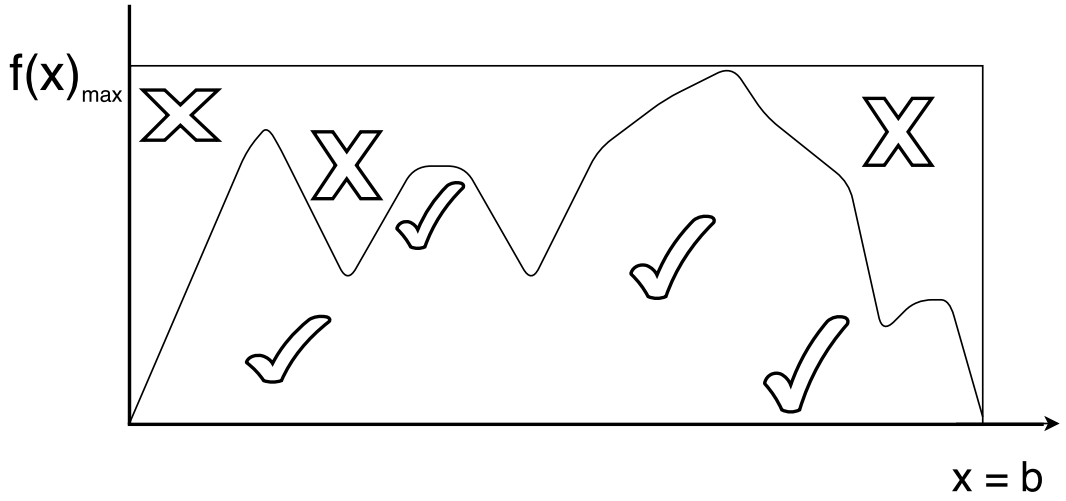


Figure 2.2: A diagramatic example of the Monte Carlo (MC) method for solving an integral. In this example, we have some function $f(x)$, and the area under the curve described by the integral $I = \int_0^b f(x)dx$. Points that will be accepted in our MC method are marked with ticks, those that are regected are marked with crosses.

2.2.4.1 Monte Carlo Radiative Transfer

As we have already seen, directly solving the equation of radiative transfer is difficult, and, at least analytically, only possible in situations with a high degree of symmetry, or homogeneity. Otherwise, the solution must be approximated, such as in Monte Carlo Radiative Transfer (MCRT). This technique aims to simulate reality by tracking photon packets through a medium, until they are absorbed or escape from the medium.

As is indicated by the name, this is a probabilistic technique, and, at its heart, relies on quantifying the probability of events that may happen to a photon, such as absorption (and in the case of radiative equilibrium, immediate re-emission), or scattering. If we consider the solution to the equation of radiative transfer,

with no source function, we have

$$I_\nu(\tau_\nu) = I_\nu(0)e^{-\tau_\nu}, \tag{2.94}$$

and from this, we can see that the probability of a photon travelling an optical depth τ_ν without being either being scattered or absorbed is

$$P(\tau_\nu) = e^{-\tau_\nu}. \tag{2.95}$$

The scattering probability is characterised by the albedo of the medium, which is simply the probability that a photon is scattered, and is given by

$$a = \frac{n_s\sigma_s}{n_s\sigma_s + n_a\sigma_a}, \tag{2.96}$$

where n is the number density, σ is the cross section, and the subscripts a and s stand for absorbed and scattered respectively. The angular phase function, $P(\Theta)$, of the scattering particle governs the probability that a photon will be scattered through an angle Θ relative to its initial direction, and has different forms, depending upon the scenario under consideration. A simple example is the isotropic phase function,

$$P(\cos\Theta) = \frac{1}{2}. \tag{2.97}$$

The basic idea, in a MCRT algorithm, is to emit N photon packets from a source (for example, a star, or a protostellar disc), and calculate the optical depth the packets experience as they move through the medium. We then compare this

calculated optical depth with a randomly sampled value of optical depth, drawn from equation 2.95. Once $\tau_{\text{calculated}} > \tau_{\text{random}}$, either scatter or absorb the photon based on the local albedo of the medium. The direction of the scattering is randomly sampled from the probabilistic phase function, and if the photon is absorbed, then *if* we are considering radiative equilibrium, the temperature of the cell is raised, and the photon is immediately re-emitted with a new frequency ν, based upon the local emissivity of the medium. By doing so, energy is conserved, since the total intensity of the packet is unchanged. If we are not considering radiative equilibrium, then energy is not (necessarily) conserved, and so the total intensity of the packet may be reduced. This process is repeated until the photon packet is either completely absorbed, or has exited the medium. At this stage, it is possible to produce an image, by creating a pixelated plane (analogous to the CCD of a camera), any photons that are incident on this plane have their position and intensity marked.

Note that this method does not allow for the additional emission of photons in the medium. For example, if a photon packet of frequency ν is absorbed and re-emitted with a lower frequency ν_2, although emitting two photons with frequency $\nu_3 = 0.5\nu_2$ each would conserve energy, the calculation proceeds until all initial photon packets, the total number of which is determined by the user, have either escaped the medium or been completely absorbed. Allowing the creation of additional photon packets would result in the calculation continuing indefinitely. Furthermore, these photon packets are meant to represent a bundle of photons of a given total energy, rather than individual photons. In this way, we can subsume interactions that result in multiple emissions into the emission

of this single, averaged, photon packet.

The number of photon packets selected by the user is important. With increasing number, the temperature structure of the medium will tend towards a converged solution. Since increasing the number comes with increased computational cost, it is best to aim for the minimum number of photon packets that will give this converged answer. For a typical hydrodynamical simulation of a protostellar disc, $\sim 10^8 - 10^9$ photon packets will give a converged answer with a significantly high signal-to-noise ratio to have a good quality image on the pixelated plane.

2.3 Protostellar Disc Structure

Stars form from a collapsing region of a molecular cloud, and since these molecular clouds are large, even though they are diffuse, even a small velocity anisotropy results in a large amount of angular momentum. This angular momentum is too large to allow the material to directly collapse onto the young star, and results in a protostellar disc due to angular momentum conservation. For the first few million years after their formation, stars host these protostellar discs, composed of dusty gas, in which planets may eventually form.

For the disc to accrete onto the central star, this angular momentum must be redistributed outwards. This process is slow, much slower than the dynamical timescale of the disc, and so we can often approximate protostellar discs as *static* objects. This is a useful approximation, that allows us to study the density and temperature profiles of these discs, without the additional complication of time. We discuss the evolution of the discs, with time, in section 2.4.

2.3.1 Classification of Young Stellar Objects

At the time of writing, we have entered "the ALMA era[5]" of observational astronomy. Like never before, we are able to directly image protostellar discs, such as HL Tau (ALMA Partnership et al., 2015), TW Hydrae (Andrews et al., 2016), Elias 2-27 (Pérez et al., 2016) and, for the first time, a disc imaged in the process of fragmentation[6]: L1448 IRS 3B (Tobin et al., 2016). These magnificent images are shown to the reader in Figure 2.3, which depicts (clockwise from top left) HL Tau, TW Hydrae, Elias 2-27 and L1448 IRS 3B. Note that the last image, L1448 IRS 3B has captured, for the first time, a protostellar disc undergoing fragmentation.

However, this capability is recent; there are, currently, only a handful of systems that have been imaged in this much detail. The TW Hydrae association, ~ 50 pc away in the constellation of Hydra, has around 30 stars, few of which have been studied in much detail. To statistically constrain characteristics of protostellar discs requires a large (or at least statistically significant) sample size, so we must go further, to Ophiuchus (~ 120 pc), Taurus (~ 150 pc) and Orion (~ 410 pc) to satisfy this requirement. At larger distances, it is more difficult to resolve these discs, but their unresolved spectral energy distributions (SEDs) can provide us

[5]ALMA stands for the Atacama Large Millimeter/submillimeter Array, located on the Chajnantor plateau in northern Chile. Comprised of 66 antennas, it is built on a partnership of eight nations, and is the largest astronomical project in history. More details can be found on the ALMA website.

[6]This is discussed in detail in section 2.5.2, but the basic idea is that regions of the disc collapse faster than they can collide with another region to be reheated. So long as this process is sufficiently rapid, these dense regions with not be sheared away, ultimately forming gravitationally bound objects.

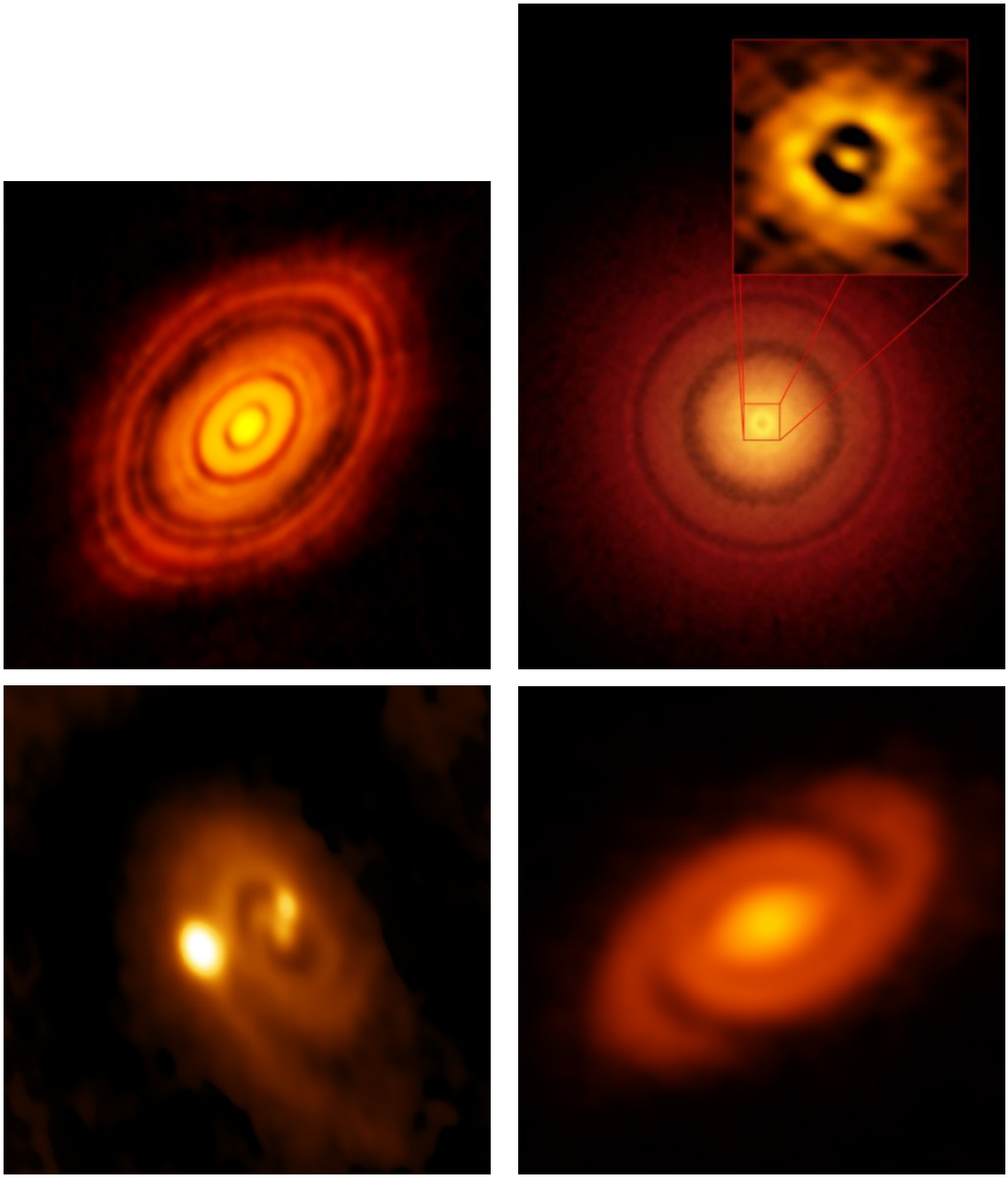


Figure 2.3: Clockwise from top left: HL TAU (credit:ALMA ESO/NAOJ/NRAO), TW Hydrae (credit:S. Andrews (Harvard-Smithsonian CfA)/ALMA ESO/NAOJ/NRAO), Elias 2-27 (credit: B. Saxton (NRAO/AUI/NSF); ALMA ESO/NAOJ/NRAO), L1448 IRS 3B (credit: B. Saxton (NRAO/AUI/NSF); ALMA ESO/NAOJ/NRAO). HL Tau and TW Hydrae show concentric ring structure, TW Hydrae, being much closer to Earth than HL Tau, has a smaller angular resolution. The inner 1 au of TW Hydrae contains a gap, which may indicate the formation of an Earth-like planet. Elias 2-27 shows two large, grand-design type spirals, typical signatures of gravitational instability within a massive disc. L1448 IRS 3B shows the collapse of a protostellar disc into three bound objects, much more massive than planets.

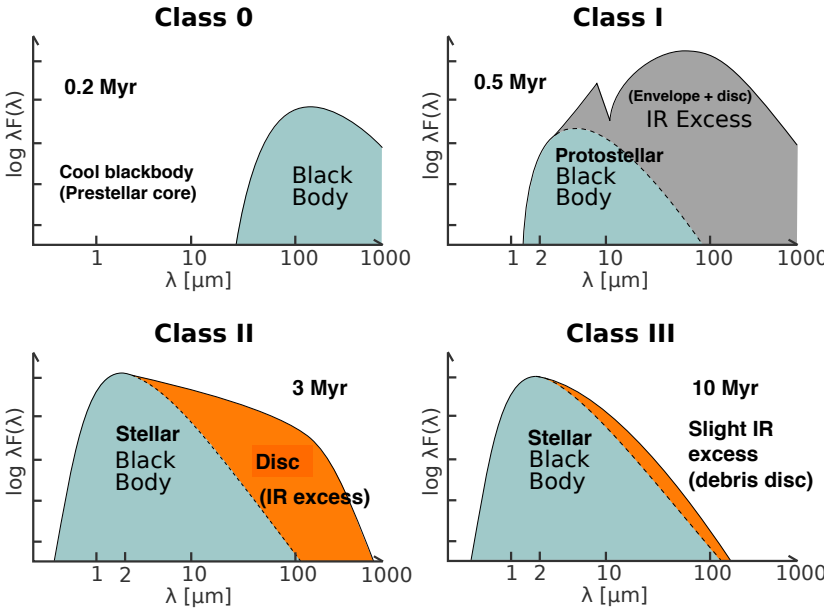


Figure 2.4: SEDs of different classes of protostar. We begin at Class 0, with a cool blackbody prestellar core, the evolves to a Class I system with a significant IR excess due to light from the protostar being reprocessed in the disc and surrounding envelope. As the system moves to the Class II phase, the central star is now a stellar blackbody, still with a significant amount of IR excess due to stellar light being processed by the disc. Finally, the system moves to the Class III phase, with a very small IR excess due to a sparse debris disc. Image credit: Adapted by C Hall from an original image by Magnus Vilhelm Persson[7].

with derived information, such as frequency and lifetime.

It is these SEDs that defined the classification system of young stellar objects (YSOs). By convention, YSOs are classified based on the *slope* of the SED, defined as

$$\alpha \equiv \frac{\delta \log(\lambda F_\lambda)}{\delta \lambda}, \tag{2.98}$$

where λ is wavelength and F is flux. This is measured in the near infra-red (IR), typically between 2.2 μm and 24 μm. A typical classification scheme is shown in Figure 2.4, and may be summarised as follows:

- **Class 0**: No flux in $\sim$ NIR ($\lambda \lesssim 10$ μm), peak of SED is in far-IR or mm. Such an object is heavily embedded in optically thick dusty gas, this stage represents the earliest part of stellar formation, possibly before the object is rotationally supported.

- **Class I**: $\alpha_{\mathrm{IR}} \geq -0.3$, still embedded in an infalling envelope, at this stage discs may first be detected, but distinguishing between disc and envelope emission may require molecular line observations. The envelope reprocesses stellar and disc radiation to longer wavelengths, and outflows and jets are common. The SED is broader than the stellar blackbody due to reprocessing of light in the disc and envelope, with λF_λ increasing beyond 2 μm.

- **Class II**: $-1.6 \leq \alpha_{\mathrm{IR}} < -0.3$, envelope has now mostly been accreted, with an SED that is still broader than blackbody, this time due only to reprocessing of light in the protostellar disc, with λF_λ either constant or decreasing beyond 2 μm. SEDs are typically well fitted by modelling the sum of stellar radiation (which is now visible at optical wavelengths, thanks to the lack of envelope), and IR and mm from the protostellar disc.

- **Class III**: $\alpha_{\mathrm{IR}} < -1.6$, the SED now matches that of a stellar blackbody, if a slight IR excess is present, this is evidence of a debris disc. Even without this IR excess, a YSO will occupy a space above the main sequence (MS) on a Hertzsprung-Russel (HR) diagram (since they are cooler for a given

[7]Freely available from the image sharing website www.figshare.com. DOI: https://doi.org/10.6084/m9.figshare.1121574.v2

luminosity)[8].

2.3.2 The Thin Disc and Low Mass Approximation

For the rest of this thesis, we have assumed that any protostellar disc we are considering is *geometrically thin*, such that the disc scale height, H, is much less than its radial extent, R, or, more properly, the *aspect ratio* is

$$\frac{H}{R} \ll 1. \tag{2.99}$$

This is important, since it allows us to make simplifying assumptions that ensure the disc may be treated as Keplerian. If this breaks down, such that $H \sim R$, which may be the case for a radiatively inefficient disc, then neither the radial pressure gradient, nor radial heat flow may be neglected (Pringle, 1981). The disc would then no longer be Keplerian, instead behaving as a centrally condensed star.

A further simplification that is made in the equations of protostellar disc structure is that the mass of the disc, relative to the mass of the central star, is low, such that $M_{\rm D} \ll M_*$. This allows us to neglect the contribution from the disc mass to the gravitational potential. This generally becomes problematic for discs when $M_{\rm D} > 0.1\ M_*$, since the gravitational potential from the disc is no longer negligible.

We will discuss this more in chapter 6, but one of the largest problems in disc theory is that an analytical description for what are known as *global* discs is less

[8] A HR diagram plots the luminosity (or absolute magnitude) of stars as a function of their temperature (or (B-V) colour difference). They are particularly useful for understanding the evolutionary path of stars

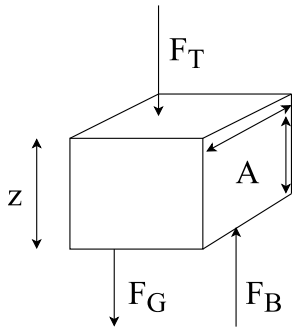


Figure 2.5: Forces on a cubic element of fluid, of height z and face area A, in hydrostatic equilibrium. F_{T} is the pressure force exerted on the top of the fluid element from the material above it, F_{B} is the pressure force on the bottom of the fluid element exerted by the material below it, and F_{G} is the force on the element due to gravity.

tractable than for *local* discs. In the global scenario, low wavenumber spirals exert global torques on the disc, so that the local approximation poorly describes the behaviour and evolution of these discs. Since low wavenumbers dominate in more massive discs (Dong et al., 2015b; Lodato & Rice, 2005), this is a problem for discs when the disc mass is a significant fraction of the central star mass.

2.3.3 Vertical Structure

Concerning ourselves for now only with a steady state solution, we can describe the vertical structure of a disc by proceeding as follows. We begin with a cube element of fluid with volume V, side area A and side length z as shown in Figure 2.5. There are three forces acting on this element, the force due to gravity, F_{G}, and the two pressure forces, F_{B} and F_{T}. To remain at rest, these three forces must sum to zero. We then have

$$F_{\mathrm{T}} + F_{\mathrm{B}} + F_{\mathrm{G}} = 0. \tag{2.100}$$

Using that this force due to pressure, $F = PA$, and that the gravitational force is $F = mg = g\rho V = g\rho zA$, where g is acceleration due to gravity, we then have

$$-P_{\mathrm{T}}A + P_{\mathrm{B}}A - \rho gzA = 0. \tag{2.101}$$

We then divide through by A, and if z is small enough to be considered infinitesimal, then we can replace z with dz. If the pressure difference, $P_{\mathrm{B}} - P_{\mathrm{T}}$, is also small enough to be considered infinitesimal, we then have

$$dP = \rho g dz. \tag{2.102}$$

The acceleration due to a gravitational potential, $\mathbf{\Phi}$, is given by $\mathbf{g} = -\nabla\mathbf{\Phi}$. Since we are only considering the vertical motion here, we only require the z-component, this leaves us with

$$\frac{dP}{dz} = -\rho\frac{\partial\mathbf{\Phi}}{\partial z}, \tag{2.103}$$

i.e. the disc is in hydrostatic equilibrium. Since we have assumed that the disc mass is sufficiently low so that it contributes little to the gravitational potential, then we have, for the gravitational potential,

$$\mathbf{\Phi} = -\frac{\mathrm{G}M_*}{r}, \tag{2.104}$$

where M_* is the mass of the central star, and $r = \sqrt{R^2 + z^2}$. Substituting this into equation 2.103, we have

$$\frac{1}{\rho}\frac{dP}{dz} = \frac{\partial}{\partial z}\left(\frac{\mathrm{G}M_*}{(R^2 + z^2)^{\frac{1}{2}}}\right). \tag{2.105}$$

Now, making use of the thin disc approximation ($z \ll R$), we have

$$\frac{1}{\rho}\frac{dP}{dz} = -\frac{\mathrm{G}M_* z}{R^3} = \Omega^2 z, \tag{2.106}$$

where $\Omega = \sqrt{(\frac{\mathrm{G}M}{r^3})}$ is the Keplerian orbital frequency. If we assume that the gas can be described by a barotropic equation of state, then we have

$$P = P(\rho),\ c_s^2 = \frac{dP}{d\rho}, \tag{2.107}$$

where c_s is the sound speed of the medium. By the relation $\frac{\partial P}{\partial z} = \frac{\partial P}{\partial \rho} = \frac{\partial \rho}{\partial z}$, we can rewrite equation 2.106 to give

$$\frac{\partial \rho}{\partial z} = -\frac{\mathrm{G}Mz}{c_s^2 r^3}\rho. = -\frac{\Omega^2 z}{c_s^2}\rho. \tag{2.108}$$

Now, we have[9]

$$\frac{1}{\rho}\frac{d\rho}{dz} = \frac{d\log\rho}{dz} = -\frac{\Omega^2}{c_s^2}z. \tag{2.109}$$

Integrating between $z = 0$ (at which $\rho = \rho_0$), we arrive at

$$\rho(z) = \rho_0 \exp\left(-\frac{\Omega^2 z^2}{2c_s^2}\right). \tag{2.110}$$

[9]Here, we have made use of the change of basis for logarithms, given by $\log_a x = \frac{\ln x}{\ln a}$, differentiating gives $\frac{d}{dx} = \frac{1}{\ln a}\frac{d}{dx}(\ln x) = \frac{1}{x \ln a}$, then using the chain rule, $f(g(x))' = g'(x)f'(g(x))$, and selecting a basis of e, we have $\frac{d}{dx}\log_e[f(x)] = \frac{1}{f(x)ln(e)}\frac{df(x)}{dx}$.

From this, we can see that the *scale height* in the disc, i.e. the length over which the density will decrease by a factor of e, is given by

$$H = \frac{c_s}{\Omega}. \tag{2.111}$$

We note that equation 2.110 shows that vertical structure of a disc is a Gaussian distribution, where H is the standard deviation, so the mass distribution is 0.68, 0.95 and 0.997 of the mass within H, $2H$ and $3H$ respectively. For the case of a self-gravitating disc, the vertical density is described by (Spitzer, 1942)

$$\rho = \frac{\rho_0}{\cosh^2(\frac{z}{H_{\rm sg}})}, \tag{2.112}$$

where

$$H_{\rm sg} = \frac{c_{\rm s}^2}{\pi {\rm G} \Sigma}. \tag{2.113}$$

2.3.4 Radial Structure

It is not possible to determine the density profile of an accretion disk without either the use of observations, or by considering the nature of angular momentum transport (which requires some sort of viscosity; we discuss this in section 2.4.2). However, we can derive the orbital velocity of the gas despite this, and so we proceed. In the thin disc approximation, the radial pressure gradient may be ignored.

Assuming, for now, that the disc is inviscid, then from equation 2.38 we have

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P - \nabla \mathbf{\Phi}, \tag{2.114}$$

where we have replaced the body force component with our expression for gravitational force. In cylindrical polar coordinates, considering only the radial components, we then have

$$\frac{\partial v_r}{\partial t} + v_r \frac{\partial v_r}{\partial r} + \frac{v_\phi}{r}\left(\frac{\partial v_r}{\partial \phi}\right) + v_z \frac{\partial v_r}{\partial z} = -\frac{1}{\rho}\frac{\partial P}{\partial r} - \frac{\partial \mathbf{\Phi}}{\partial r}. \tag{2.115}$$

If v_r is sufficiently small, we can discount the first two terms in equation 2.115. For an axisymmetric disc, $\frac{\partial v_r}{\partial \phi} = 0$, and discounting, the pressure gradient due to the thin disc approximation, all we are left with is

$$\frac{v_\phi^2}{r} = \frac{\partial \Phi}{\partial r}, \tag{2.116}$$

indicating a centrifugal balance in the disc. For a non self-gravitating disc, $\mathbf{\Phi} = -\mathrm{G}M_*/r$, then we can determine the Keplerian orbital velocity,

$$v_\phi = v_\mathrm{K} = \sqrt{\frac{\mathrm{G}M_*}{r}} \tag{2.117}$$

and thus the Keplerian orbital frequency is

$$\Omega = \frac{v_\mathrm{K}}{r} = \sqrt{\frac{\mathrm{G}M_*}{r^3}}. \tag{2.118}$$

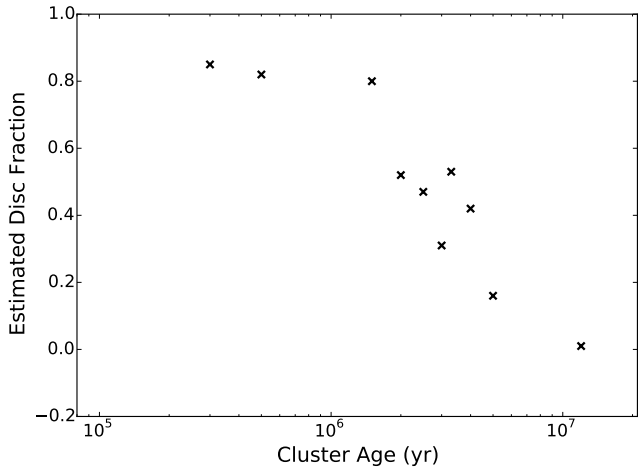


Figure 2.6: Disc fraction estimate as a function of cluster age. Error bars are not shown for clarity, data taken from Haisch et al. (2001a), Gutermuth et al. (2008), Hernández et al. (2007), Lada et al. (2006), Sicilia-Aguilar et al. (2006) and Damjanov et al. (2007). The decreasing disc fraction with cluster age provides observational evidence for the evolution of protostellar discs.

2.4 Protostellar Disc Evolution

Protostellar discs are not static, they evolve slowly over time. The observational evidence for this is rooted in the fact that disc signatures, such as IR excess, or flux at mm wavelength, is only located around stars that are in (or near) star forming complexes. Furthermore, surveys that estimate the disc fraction (the number of stars with discs compared to all the stars in the cluster) in young clusters have found a strong decline with age. An illustrative example is shown in Figure 2.6, which is a compilation from several surveys.

Explaining, theoretically, why - or how - protostellar discs evolve is not easy. Since discs contain angular momentum, for material to travel inwards onto the central star, angular momentum must be transferred outwards. This is *the* central problem in accretion disc theory, since it is a problem that accretion discs around

black holes and other compact objects must overcome.

In this section, we discuss the time evolution of the surface density of protostellar discs, and discuss our traditional understanding of turbulent viscosity. We touch briefly on several sources of turbulence, but focus predominantly on gravitational instability, since that is the focus of this thesis.

2.4.1 Surface Density Evolution

To construct equations for the evolution of a protostellar disc, we follow Pringle (1981) and begin with the derivation of a continuity equation for the disc. Consider an azimuthally symmetric annulus, with radius R, surface density Σ and radial extent ΔR. Such an annulus has mass m given by

$$m = 2\pi R \Delta R \Sigma. \tag{2.119}$$

The rate of change of this mass is then

$$\frac{\partial m}{\partial t} = \frac{\partial}{\partial t}(2\pi R \Delta R \Sigma), \tag{2.120}$$

which, making use of the product rule, and realising that $\frac{\partial R}{\partial t} = 0$, can also be written as

$$\frac{\partial m}{\partial t} = 2\pi R \Delta R \frac{\partial \Sigma}{\partial t} \tag{2.121}$$

The rate of change of mass in the annulus is determined by the flow of mass into and out of the annulus, with a velocity $v_r(R)$ at the R boundary, and $v_{\mathrm{R}}(R+\Delta R)$ at the $R + \Delta R$ boundary. Note that only the radial component of this velocity

is important. This can be stated as

$$\frac{\partial}{\partial t}(2\pi R\Delta R\Sigma) = 2\pi R\Sigma v_{\rm R}(R) - 2\pi(R+\Delta R)\Sigma(R+\Delta R)v_{\rm R}(R+\Delta R). \quad (2.122)$$

If we now divide both sides by $2\pi\Delta R$, and make use of equation 2.121, we then have

$$R\frac{\partial\Sigma}{\partial t} = \frac{Rv_{\rm R}(R)\Sigma(R) - (R+\Delta R)v_{\rm R}(R+\Delta R)\Sigma(R+\Delta R)}{\Delta R} \quad (2.123)$$

So long as ΔR is small, then in the limit $\Delta R \to 0$, we have

$$R\frac{\partial\Sigma}{\partial t} + \frac{\partial}{\partial R}(R\Sigma v_r) = 0. \quad (2.124)$$

We now proceed in a similar fashion with the conservation of angular momentum across the annulus. The angular momentum of an annulus is given by

$$L = mv_\phi R, \quad (2.125)$$

and using our expression for mass in an annulus, coupled with $v_\phi = \Omega R$, we have

$$L = 2\pi\Omega R^3\Delta R\Sigma, \quad (2.126)$$

the time derivative of this is

$$\frac{\partial L}{\partial t} = 2\pi R\Delta R\frac{\partial}{\partial t}(\Sigma\Omega R^2). \quad (2.127)$$

We leave the equation in this form, since on the right hand side, $2\pi R\Delta R$ is the area of the annulus, and $\Sigma\Omega R^2$ is the angular momentum per unit area. Considering, as with mass, the flow of angular momentum in and out of the annuli, we have the contribution *in* from R, *out* through $R+\Delta R$, and the viscous torque, Γ, one contribution from the inner annuli, rotating more rapidly, and one contribution from the outer annuli, rotating more slowly. Combined, we have

$$\frac{\partial L}{\partial t} = L_{\rm in}(R) - L_{\rm out}(R+\Delta R) + \Gamma_{\rm in}(R) - \Gamma_{\rm out}(R+\Delta R). \tag{2.128}$$

Now, as before with mass, the angular momentum is transported at radial velocity v_r. We now have

$$\begin{aligned}\frac{\partial L}{\partial t} &= 2\pi R\Sigma(R)R^2\Omega(R)v_r(R) \\ &\quad - 2\pi R\Sigma(R+\Delta R)R^2\Omega(R+\Delta R)v_r(R+\Delta R) \\ &\quad + \Gamma(R) - \Gamma(R+\Delta R).\end{aligned} \tag{2.129}$$

Equating equations 2.127 and 2.129, dividing through by $2\pi\Delta R$, and in the limit that $\Delta R \to 0$, we have

$$R\frac{\partial}{\partial t}(\Sigma\Omega R^2) + \frac{\partial}{\partial R}(R\Sigma R^2\Omega v_r) = \frac{1}{2\pi}\frac{\partial}{\partial R}\Gamma. \tag{2.130}$$

Here, Γ is the net torque across the annulus. At this point, we must, at least, attempt to consider the origin of the torque. The most simple solution is to simply assume this torque is due to ordinary kinematic viscosity, i.e. some sort of

resistance to shear stress[10]. Since we have assumed the disc is Keplerian, it will be differentially rotating, and there will be an associated shear between adjacent annuli due to the viscosity. The shear rate for two fluid sheets, is $A = \frac{\Delta v_\phi}{\Delta r}$, in the infinitesimal limit, for a Keplerian disc, this becomes

$$A = R\frac{d\Omega}{dR}, \tag{2.131}$$

where we keep the first R in the right hand term constant. The viscous force, per unit length, around the circumference of an annulus is $\nu\Sigma A$, so the torque becomes

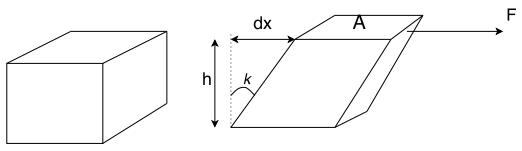
$$\Gamma = 2\pi R\,\nu\Sigma A\,R. \tag{2.132}$$

We can now use this expression in equation 2.130 to give

$$\frac{\partial}{\partial t}(\Sigma\Omega R^2) + \frac{1}{R}\frac{\partial}{\partial R}(\Sigma R^3\Omega v_r) = \frac{1}{R}\frac{\partial}{\partial R}\left(R^3\nu\Sigma\frac{d\Omega}{dR}\right). \tag{2.133}$$

Expanding the left hand side of the equation using the product rule, assuming $\frac{d\Omega}{dt} = 0$ (which is reasonable, considering the timescales over which Ω will evolve), dividing through by $\frac{\partial}{\partial R}(\Omega R^2)$, and making use of the mass continuity equation

10



Viscosity is typically defined as $\nu = G \cdot t$, where G is the shear modulus given by $G = \frac{\text{shear stress}}{\text{shear strain}} = \frac{F/A}{dx/h}$. We can see that the value of ν is large when in a given time, the amount of deformation, dx is low for a given force, F. Alternatively, a longer time is required to achieve the same level of deformation for a given force on a fluid element on an object with higher viscosity than an object with lower viscosity.

given in equation 2.124, we have

$$\frac{\partial \Sigma}{\partial t} = -\frac{1}{R}\frac{\partial}{\partial R}\left(\frac{1}{\frac{\partial}{\partial R}(\Omega R^2)} \cdot \frac{\partial}{\partial R}\left[R^3 \nu \Sigma \frac{d\Omega}{dR}\right]\right). \tag{2.134}$$

Now, we once again rely upon our assumption of Keplerian rotation, and substitute $\Omega = (\mathrm{G}M/R^3)^{1/2}$ into equation 2.134 to give us, finally, the equation describing the evolution of a viscous accretion disc in Keplerian rotation

$$\frac{\partial \Sigma}{\partial t} = \frac{3}{R}\frac{\partial}{\partial R}\left[R^{\frac{1}{2}}\frac{\partial}{\partial R}(\nu \Sigma R^{\frac{1}{2}})\right]. \tag{2.135}$$

2.4.1.1 Steady State Solution

In chapter 6, we make use of an analytical model that relies upon the steady state solution to the equation of viscous evolution of an accretion disc. To show how we arrive at the steady state solution, we begin at equation 2.133, and noting that by definition, in the steady state solution, the temporal dependence is removed, such that time derivatives are 0, we have

$$\frac{1}{R}\frac{\partial}{\partial R}(\Sigma R^3 \Omega v_r) = \frac{1}{R}\frac{\partial}{\partial R}\left(R^3 \nu \Sigma \frac{d\Omega}{dR}\right). \tag{2.136}$$

In a steady state, the accretion rate, $\dot{M}$, across the disc remains constant over dynamical timescales, which implies that (from equation 2.121)

$$\dot{M} = -2\pi \Sigma r v_r \tag{2.137}$$

is constant. We note that negative sign indicates the movement of material *towards* the central star. Rearranging equation 2.137 to $\Sigma v_r = \frac{\dot{M}}{2\pi r}$, we substitute this into equation 2.136 to obtain

$$\frac{\partial}{\partial R}\left(\frac{-R^2\Omega\dot{M}}{2\pi}\right) = \frac{\partial}{\partial R}\left(R^3\nu\Sigma\frac{d\Omega}{dR}\right). \tag{2.138}$$

Integrating between the inner boundary, $R_{\rm in}$, and R, and assuming no torque at the inner boundary (such that $\frac{d\Omega}{dR} = 0$), we now have

$$-\dot{M}R^2\Omega + \dot{M}R_{\rm in}^2\Omega_{\rm in} = 2\pi R^3\nu\Sigma\frac{d\Omega}{dR}. \tag{2.139}$$

Using $\Omega = \frac{{\rm G}M}{R^3}$, $\frac{d\Omega}{dR} = -\frac{3}{2}\sqrt{\frac{{\rm G}M}{R^5}}$, and rearranging gives us

$$\dot{M}\left[R_{\rm in}^2\left(\frac{{\rm G}M_{\rm in}}{R_{\rm in}^3}\right)^{\frac{1}{2}} - \left(\frac{{\rm G}M}{R^3}\right)^{\frac{1}{2}}\right] = -3\pi\nu\Sigma\left(\frac{{\rm G}M}{R^5}\right), \tag{2.140}$$

which simplifies to

$$\dot{M}\left[({\rm G}M_{\rm in}R_{\rm in})^{\frac{1}{2}} - ({\rm G}MR)^{\frac{1}{2}}\right] = -3\pi\nu\Sigma({\rm G}MR)^{\frac{1}{2}}, \tag{2.141}$$

then we divide by $({\rm G}MR)^{\frac{1}{2}}$ to obtain

$$\left[\left(\frac{M_{\rm in}R_{\rm in}}{MR}\right)^{\frac{1}{2}} - 1\right] = -3\pi\nu\Sigma. \tag{2.142}$$

Now, we may assume that the majority of the mass is contained in the central protostar, such that $M_{\rm in} \approx M$. If this is the case, and multiplying through by

-1, then we have

$$\dot{M}\left[1-\left(\frac{R_{\rm in}}{R}\right)^{\frac{1}{2}}\right]=3\pi\nu\Sigma. \tag{2.143}$$

In the limit that $R \gg R_{\rm in}$, and defining $\nu = \alpha c_{\rm s} H$ (we discuss this in more detail in section 2.4.2) where $c_{\rm s}$ is sound speed in the medium, α is a dimensionless parameter < 1 and H is the disc scale height, with $H = \frac{c_{\rm s}}{\Omega}$ we then have

$$\dot{M} = \frac{3\pi\alpha c_{\rm s}^2\Sigma}{\Omega}. \tag{2.144}$$

This equation forms the basis of the semi-analytical model for a self-gravitating disc in quasi steady equilibrium described in chapter 6, and we leave further discussion of this until then.

2.4.2 Viscosity

Until now, we have carefully ignored any consideration of the *origin* of the viscosity in the disc. Our first guess is simply ordinary molecular viscosity, which arises due to the expectation value of the distance travelled by a particle before it collides with another particle. This is known as the *mean free path*, λ, and is given by

$$\lambda = \frac{1}{n\sigma}, \tag{2.145}$$

where n is the number density of the gas, and σ is the cross-section for molecular collisions. The kinematic viscosity, ν, is related to λ by

$$\nu = \lambda c_{\rm s}, \tag{2.146}$$

where c_s is the sound speed in the medium. We can determine the characteristic timescale of this viscosity, since more generally, the viscosity is given by

$$\nu = l \cdot v \tag{2.147}$$

where l is the characteristic length[11] of the medium, and v is the characteristic velocity. Using $v = \frac{l}{t}$, we substitute this into equation 2.147, and rearrange, to give us an expression for characteristic timescale,

$$t = \frac{l^2}{\nu}. \tag{2.148}$$

Adopting reasonable values for a protostellar disc, $\sigma = 2{\times}10^{-15}$ cm^{-2}, $c_s = 0.5$ km s^{-1}, $n = 10^{12}$ cm^{-3} and $l = 10$ au, gives us a characteristic viscous timescale of $\sim 10^{13}$ years. Since this is ~ 1000 times the age of the universe, molecular viscosity is not the viscosity responsible for angular momentum transport in accretion discs.

This large viscous timescale is due to the small molecular viscosity. A small viscosity, however, corresponds to a large Reynolds number, defined as

$$Re \equiv \frac{v \cdot l}{\nu}, \tag{2.149}$$

where v and l are the characteristic velocity and length respectively. In laminar

[11]The concept of characteristic length and velocity can be slightly subtle, but it is helpful to consider a turbulent region in a cloud. The region of this turbulence may be thought of as a characteristic length. In engineering, the width of a pipe containing some fluid will often be described as the characteristic length of the fluid. It is, in general, the minimum length scale over which the properties of the system will substantially change.

flow, viscosity dominates and Re is small. When Re is large, inertial forces dominate flow behaviour and the flow becomes highly turbulent. For protostellar discs, using $Re = c_s H/\nu$, with $H = 0.1R$ and $c_s = 0.5$ km s^{-1}, at 10 au yields $Re = 3 \times 10^{10}$. At such a high number, astrophysicists generally conclude that the flow is turbulent. However, the critical number at which a flow becomes turbulent is *heavily* geometry dependent (for pipes, $Re_{\rm crit} \sim 10^3$, for aerofoil $Re_{\rm crit} \sim 10^5$), and, additionally, there is currently no conclusive observational evidence that suggests protostellar discs are, necessarily, turbulent. However, CO(3-2) observations of HD 163296 and TW Hya do yield derived linewidths consistent with turbulent broadening (Hughes et al., 2011), implying a turbulent viscosity parameter of $\alpha \sim 0.01$ in these discs. We proceed as though the disc *is* turbulent, for which the characteristic length of the turbulence is H, and the characteristic velocity of the turbulence is c_s. By using the Shakura & Sunyaev (1973) α parameter, we can write

$$\nu = \alpha c_s H, \tag{2.150}$$

where α is a dimensionless parameter, < 1, that characterises the strength of the turbulence. This is an incredibly useful prescription, since it encapsulates all uncertainty into one quantity.

2.4.3 Angular Momentum Transport and the Gravitational Instability

Having established that the entire theoretical framework of accretion disc physics relies on the notion of viscosity, and that the only possible source of this is some sort of turbulence, since molecular viscosity is far too small, we must now consider the origin of this turbulence.

Given the interest of this thesis, we do not discuss every available option in detail, we merely list a few, and where they do (and do not) become important. We do, however, cover the gravitational instability in some detail.

2.4.3.1 Hydrodynamic Instability and the Rayleigh Criterion

We first consider hydrodynamic instabilities in a rotating flow. Consider two concentric annuli, ring 1 and ring 2, at radius r_1 and r_2 with $r_2 > r_1$. The specific angular momentum of each annulus is $r^2\Omega$. Consider now a scenario where an element of fluid in these two rings is swapped. Since angular momentum must be conserved, we have that the total change in angular momentum is

$$d(r^2\Omega) = 2r\Omega\, dr + r^2\, d\Omega = 0. \tag{2.151}$$

This now gives us

$$\frac{d\Omega}{dr} = -\frac{2}{r^3}(r^2\Omega), \tag{2.152}$$

where we have kept the specific angular momentum term for clarity. Let us now consider the element of fluid from ring 1 displaced into ring 2. For stability, we

require acceleration to be restorative, acting to restore the element to its original position. This only makes sense if $\frac{d\Omega}{dr} < 0$, i.e. the acceleration is acting to restore the displaced element to its original velocity. For $\frac{d\Omega}{dr} < 0$ at all radii then implies that

$$\frac{d}{dr}(r^2\Omega) > 0. \tag{2.153}$$

This is known as the *Rayleigh criterion* for stability. Since for Keplerian discs, angular momentum increases with radius, then a disc in Keplerian rotation is predicted to never be unstable due to hydrodynamic turbulence, despite the very large Reynolds number.

2.4.3.2 Other Sources of Instability

Since ordinary hydrodynamic turbulence cannot be the source of angular momentum transport in Keplerian discs, we are forced to consider other options. The first is the magnetorotational instability (MRI). At very early times, shortly after formation, the majority of the disc is likely to be un-ionized, and so the MRI will be low. However, at later times when the ionization is high, this can become the dominant angular momentum transport in the disc. If we imagine an accretion disc that is vertically threaded with magnetic field lines, an displace an element of fluid from r_1 to r_2, the magnetic field will try to move the element back to r_1, since the field does not want to be stretched.

However, the field will also attempt to resist the shearing motion of the disc, which can lead to instability. Without the presence of a magnetic field, a fluid element that is displaced radially outwards will decrease its angular velocity, to settle into equilibrium with the rest of the flow. However, in the presence of a magnetic

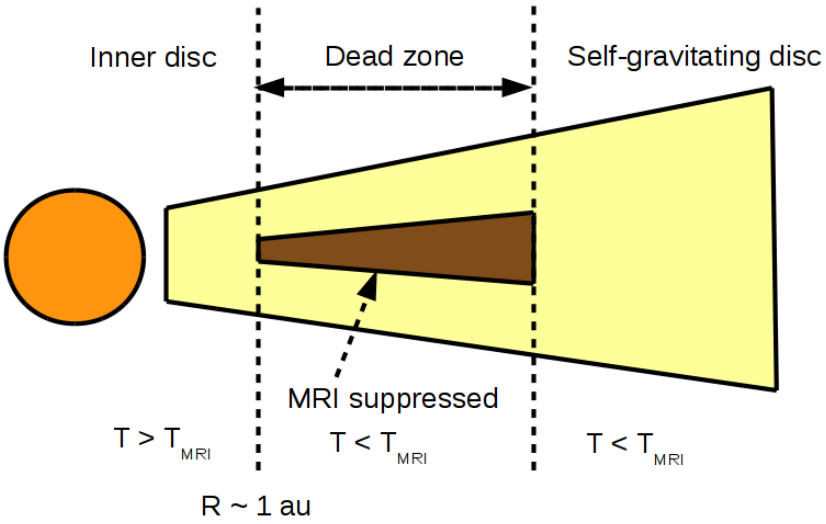


Figure 2.7: Radial structure of a layered disc (see, e.g., Armitage 2011; Armitage et al. 2001; Gammie 1996. At small radii, the temperature exceeds the minimum temperature, $T_{\rm MRI}$, needed for MRI to be active, typically $T_{\rm MRI} \sim 10^3$ K. As such, MHD turbulence provides an efficient amount of angular momentum transport. In the "dead zone", the midplane temperature is too low to support MRI, and so accretion occurs in the upper layers, which remain ionised either through x-rays from the central star, or possibly cosmic rays. At larger radii, we may expect angular momentum transport to be driven by the disc self-gravity.

field, once a fluid element is displaced, then the field resists its attempt to return to its original position. This causes outwards transport of angular momentum. This is very much a qualitative description of the MRI, since we do not consider MHD in this thesis. We refer the interested reader to Balbus & Hawley (1991) for a more in-depth discussion, but we note here that the gravitational instability, GI, can interact with MRI.

Figure 2.7 shows our current understanding of ideas originally outlined by Gammie (1996), whereby discs that transport angular momentum through the MRI develop a layered structure. In the original version, cosmic rays were the main source of non-thermal ionisation in the upper layers and outer regions of

the disc; however, we now understand that x-rays are probably the main source (Glassgold et al., 1997) of this, with cosmic rays remaining a possibility in the outer parts of the disc. In the innermost region of Figure 2.7, the temperature is sufficiently high that MRI is active, so the disc will predominantly transport angular momentum in this region through MHD turbulence. In the intermediate region, which numerical simulations have shown to begin at $R \sim 1$ au (see, for example, Bai & Goodman 2009,Terquem 2008), the mid-plane is sufficiently cool that a *dead zone* develops, suppressing MRI, so accretion occurs predominantly in the MRI active upper layers. In the outer region of the disc, we expect angular momentum transport due to GI from disc self-gravity to dominate.

It is possible that GI can act to both trigger and suppress MRI. For example, mass accumulation in the *dead zone* can lead to GI, which may trigger MRI through heating. However, it has been shown that discs in which self-gravity is sufficiently strong will suppress MRI (Lin, 2014). The basic idea behind this is that MRI operates with a wavelength λ, and this wavelength needs to fit inside the disc, such that $\lambda \lesssim 2H$ (Sano & Miyama, 1999). For a self-gravitating disc, the self-gravitating scale height, $H_{\rm SG}$, is given by

$$H_{\rm SG} = \frac{c_s^2}{\pi \mathrm{G}\Sigma}, \tag{2.154}$$

from which we can see that disc thickness decreases with increasing self-gravity. This makes it more likely that the requirement $\lambda \lesssim 2H$ will not be satisfied, and, therefore, that MRI will be suppressed.

Another possibility is that complex radiative processes may cause the disc to

become thermally unstable, undergoing phase changes on large scales. This could certainly create large, turbulent eddies. However, it is to the *gravitational instability* that we now turn our attention.

2.4.3.3 Gravitational Instability

For massive discs, with disc-to-star mass ratios $q \sim 0.1$ or more, their self-gravity becomes important. Under these conditions, a perturbation will grow, and in the case of differential rotation, will shear and grow into spiral density waves. Strictly speaking, to derive the conditions under which a disc is gravitationally unstable requires us to consider the response of the disc to a small perturbation. This approach is found in Binney & Tremaine (2008), having originally been derived by Lin & Shu (1964). However, this is less intuitive than the more simple approach we take here, for the same qualitative result.

If we consider a region of a disc, with gravitational potential energy, rotational kinetic energy and thermal energy $E_{\rm G}$, $E_{\rm R}$ and $E_{\rm T}$, respectively, then for the disc to become unstable to collapse, we require

$$E_{\rm R} + E_{\rm T} < |E_{\rm G}|. \tag{2.155}$$

If the region of the disc is of size a, then the mass of the region is given by

$$M_{\rm R} = \Sigma a^2. \tag{2.156}$$

The gravitational potential energy is then

$$E_{\mathrm{G}} = -\frac{\mathrm{G}M_{\mathrm{R}}^2}{a} = -\mathrm{G}\Sigma^2 a^3, \tag{2.157}$$

the rotational kinetic energy is

$$E_{\mathrm{R}} = \frac{1}{2}M_{\mathrm{R}}\Omega^2 a^2 = \frac{1}{2}\Sigma\Omega^2 a^4. \tag{2.158}$$

Substituting and rearranging, we find

$$\Sigma a^2(-2\mathrm{G}\Sigma a + \Omega^2 a^2 + c_{\mathrm{s}}^2) < 0. \tag{2.159}$$

Since the bracketed term is quadratic in a[12], we have a minimum at

$$a = \frac{\mathrm{G}\Sigma}{\Omega^2}. \tag{2.160}$$

Substituting this into the condition for instability given in equation 2.159, we then find

$$\frac{c_{\mathrm{s}}\Omega}{\mathrm{G}\Sigma} < 1. \tag{2.161}$$

This is, approximately, the Toomre (1964) parameter for gravitational instability. More properly, it is given by

$$Q = \frac{c_{\mathrm{s}}\Omega}{\pi\mathrm{G}\Sigma} < 1. \tag{2.162}$$

[12]The x-coordinate of a minima in a quadratic expression such as $y = ax^2 + bx + c$ is found by differentiating the expression, and setting this equal to 0.

The full derivation for this, beginning from the dispersion relation in the tightly-wound (WKB) approximation, can be found in appendix A. We can see that increasing the temperature or rate of rotation will cause Q to rise, making the disc more stable to gravitational instability. Increasing the mass of the disc will cause Q to fall, making the disc susceptible to gravitational instability. The condition that $Q < 1$ is the condition for the disc to become unstable to axisymmetric modes. For non-axisymmetric modes, it has been shown in the review by Durisen et al. (2007) that $Q < 1.5 - 1.7$

Through the Toomre parameter, we can understand how a self-gravitating disc may settle into a self-regulating state of marginal instability. When a disc is hot, such that $Q \gg 1$, it radiatively cools, which decreases Q. As Q decreases, and approaches $Q \sim 1.5 - 1.7$ (Durisen et al., 2007), it becomes susceptible to non-axisymmetric perturbations and spiral waves develop. More rapid cooling leads to large spiral amplitudes, and larger spiral amplitudes heat the disc more rapidly through local shock heating. This heating increases the sound speed, driving the Q value up until the disc is no longer unstable, and generation of spiral waves will then cease. This state, where Q remains $Q \sim 2$, is known as *the self-regulated state of marginal instability* (Paczynski, 1978).

The steady production of spirals in this marginally unstable state is the cause of gravito-turbulence in the disc. So long as this is present, this turbulence will provide an effective viscosity, allowing us to describe our hydrodynamical discs using the viscous parameterisation. However, we note here that the viscous parameterisation is only valid for the *local approximation*. The local approximation is discussed in more detail in Appendix A, but the basic idea

is that for angular momentum transport to occur locally, then spirals must be sufficiently tightly wound. If the spirals are loose, as may be seen in simulations of more massive discs, then the long-range gravitational force may no longer be neglected. The local approximation will no longer be valid, and thus the evolution of our discs using the viscous parameterisation will not strictly be valid.

2.5 Planet Formation

The definition of a "planet" sprang into public consciousness in 2006, when the International Astronomical Union (IAU) downgraded Pluto from "planet" to "dwarf planet", since it has not cleared its orbit of other bodies. At the other end of the scale, we must consider a cut-off point, where a planet is not really a "planet" any more. With this in mind, we now define some criteria that a "planet", according to the IAU, must meet:

- In orbit around the Sun.
- Is in hydrostatic equilibrium.
- Has cleared the neighbourhood around its orbit.

However, this was constructed with our own solar system in mind, and explicitly excludes exoplanets, so does not address the issue of an upper mass limit. At ~ 13 M_J, a "planet" will begin to fuse deuterium in its core, any object that does this is known as a *brown dwarf*. We therefore include that any "planet" must also not fuse deuterium in its core, and so has an upper mass limit of 13 M_J.

Any theory of planet formation must be able to explain the menagerie of

exoplanetary configurations that is now known to us. We know that "hot Jupiters", $\sim$ Jupiter mass planets on a Mercury-like orbit, occur around $\sim 1.2\%$ of stars (Wright et al., 2012), and that free-floating, sub-stellar objects, such as WISEA 1147-2040 and 2MASS 1119-1137 exist, so any planet formation theory must be able to explain their incidence.

At present, there are two complementary theories of planet formation: core accretion and gravitational instability. We discuss both of these now, but we highlight the fact that it has been known for some time that gravitational instability in protostellar discs rarely forms planetary mass objects (see, for example, Rice et al. 2015). Nevertheless, the core accretion theory is faced with several challenges, which may or may not be solved by the addition of processes such as the streaming instability, so we include a discussion of planet formation through gravitational instability here.

2.5.1 Core Accretion

In the core accretion scenario of planet formation, there are three different scales that are dominated by different physical processes. First is the formation of planetesimals, objects that are large enough ($\sim$ 10 km) that their evolution is dominated by gravitational interactions rather than by interaction with the gas flow. To begin, we consider a fluid disc of dusty gas. The gas will orbit at just below Keplerian velocity, with this velocity difference originating from the radial pressure gradient exerted on the gas.

To see this, we begin with equation 3.90, which is a statement of centrifugal

balance in the disc. This time, we do not neglect the partial pressure support in the disc, so that equation 3.90 is now

$$\frac{v_\phi^2}{r} = \frac{\partial \Phi}{\partial r} + \frac{1}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r}. \tag{2.163}$$

We substitute $\Phi = -\mathrm{G}M_*/r$ into equation 2.163 to retrieve

$$\frac{v_\phi^2}{r} = \frac{\mathrm{G}M_*}{r^2} + \frac{1}{\rho}\frac{\mathrm{d}P}{\mathrm{d}r}. \tag{2.164}$$

Since the Keplerian velocity is $v_\mathrm{K} = (\mathrm{G}M_*/r)^{\frac{1}{2}}$, we can now see that v_ϕ will be less than v_K for the gas. This will become important when we consider solid bodies orbiting in the gas, since *they do not feel this pressure gradient*. Assuming that the pressure-law is some power law function of radius, we have

$$P = P_0 R^{-n}, \tag{2.165}$$

where P can be described by a locally isothermal equation of state as

$$P = \rho_\mathrm{g} c_s^2 \tag{2.166}$$

where ρ_g is gas density and c_s is sound speed. Using equation 2.165 and equation 2.166, we can now see that the pressure gradient term can be written as

$$\frac{1}{\rho_\mathrm{g}}\frac{\mathrm{d}P}{\mathrm{d}r} = -n\frac{c_s^2}{R}. \tag{2.167}$$

Subsituting this into equation 2.164, we now see that the gas orbits at a sub-

Keplerian velocity given by

$$v_\phi = v_\mathrm{K}\left(1 - n\frac{c_s^2}{v_\mathrm{K}}\right)^{\frac{1}{2}}. \tag{2.168}$$

Given some typical values for a protostellar disc, i.e. assuming $H/R = 0.05$ and $\Sigma \propto R^{-1}$ yields n=3, for which

$$v_\phi = 0.996 v_\mathrm{K}. \tag{2.169}$$

This difference is negligible when considering only the motion of gas. However, since dust particles, as solid bodies, do not feel this pressure gradient, they orbit at Keplerian velocity. This means the dust particles, which, in the absence of gas, would orbit at v_K, feel a headwind from the slower-moving gas orbitting at v_ϕ, and therefore the dust orbits at sub-Keplerian velocity. These sub-Keplerian speeds result in a centrifugal force that is not high enough to balance the gravity from the central star, and so these particles experience a radial drift inwards.

Figure 2.8 shows the radial, transverse and escape velocity for dust particles as function of particle radius, taken from Weidenschilling (1980). We can see that for particles of size ~ 100 cm, at a distance of 1 au, the particle would drift inwards in a timescale of just $t \sim \frac{distance}{velocity} \sim \frac{10^{13}\,\mathrm{cm}}{10^4\,\mathrm{cm\,s^{-1}}} \sim 30$ years! This particular problem is known as the *metre barrier*, because these grains have such high drift velocities, as grains reach these size they should rapidly migrate inwards and be destroyed due to their large velocities, relative to the rest of the particulate matter in the disc. This tells us that planetesimal formation needs to be rapid, otherwise the dust particles will be photoevaporated by the central star as they migrate inwards. An

example solution to the problem of rapid inward migration may be dust trapping at local pressure maxima in spiral arms, such as that described by Rice et al. (2004).

Although there is no universally accepted consensus as to how dust grains grow to form $\sim$ km sized objects, the most simple model is to assume that growth occurs through a series of pairwise interactions that on average result in growth. However, as just discussed, the large radial drift velocities of particles of size ~ 1 m means that the most likely outcome of interaction is fragmentation, rather than coagulation. A proposed solution to the metre barrier problem is the *pebble accretion theory* (Lambrechts & Johansen, 2012; Levison et al., 2015). Pebbles are grouped together due to the streaming instability (Youdin & Goodman, 2005), whereby pebbles orbit at Keplerian velocity, but the gas is pressure supported from the host stellar radiation, causing the gas to orbit at sub-Keplerian speeds. Feeling a headwind, pebbles slow, and as more pebbles migrate inwards, they group together, as the headwind reduces locally (analogous to track cyclists in their teammates' slipstream). The clusters of pebbles exert a backreaction on the gas, causing it to slow further. This reduces the local drag, and reduces the radial drift rate. When sufficiently large, these groups of pebbles gravitationally collapse (Youdin, 2011), and can begin to accrete pebbles until they form giant planet cores.

For now, let us proceed as if planetesimals definitely form. This leads to the second size scale in core accretion, that of terrestrial planet formation. Once a population of planetesimals has formed in the disc, they undergo mutual gravitational interactions, continuing to coalesce, and experiencing a

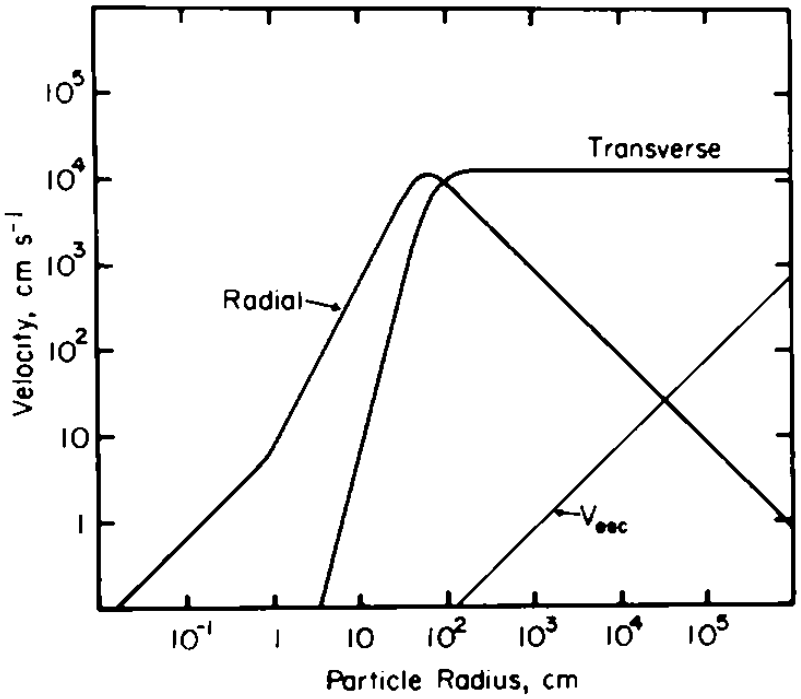


Figure 2.8: Taken from Weidenschilling (1980), Figure 1. Radial and transverse velocity at $R = 1$ au, $z = 0$ of dust particles in the solar nebula model. The radial velocity peaks at dust particles sizes of 1 m, at 1 au, since this corresponds to a Stokes[13] number of ~ 1.

small amount of hydrodynamic drag from the remaining gas, which serves to dampen the growth of eccentricity and inclination. The first phase in terrestrial planet formation is *runaway growth*, which occurs once the escape velocity of a planetesimal becomes significantly larger than its relative velocity. The growth rate is set by the mass of the planetesimal, and these planetesimals become protoplanets. The most massive protoplanets grow the most rapidly in this regime, and for this reason it is known as the oligarchic growth phase. The effect is local at this stage, rather than global across the disc, so many oligarchs will exist, feeding from their own local pool of planetesimals.

We then enter the third and final stage of planet formation via core accretion, the giant planet formation and core migration phase. At around ~ 1 $M_\oplus$, the protoplanet becomes gravitationally coupled to the gas in the disc, resulting in eccentricity damping and exchange of angular momentum. This exchange of angular momentum can cause the planet to move inwards, known as *migration*. A sufficiently massive protoplanet (~ 10 $M_\oplus$) is able to accrete a substantial envelope to form its atmosphere, forming a giant planet. The giant planet ceases to accrete mass when there is no longer gas available, whether that is because the disc's gas has photoevaporated, or because the forming protoplanet has accreted all of the available mass inside it's Hill sphere, which is the sphere of influence where the gravity of the forming planet dominates locally over the global gravitational potential, given by the Hill radius,

$$R_\mathrm{h} = a\left(\frac{M_\mathrm{p}}{3M_*}\right)^{\frac{1}{3}}, \tag{2.170}$$

where a is the semi-major axis of the planetesimal's orbit, and $M_{\rm p}$ is the mass of the planetesimal. At distances of ~ 5 au from the central star, Lissauer et al. (2009) have shown that it is possible to form a Jupiter mass planet within the lifetime of the disc. However, there are many wide orbit planets, of ~ 5 $\rm M_J$ or greater, that exist beyond 40 au from their host star, such as the objects directly imaged in the HR 8799 system (Marois et al., 2008). There is simply not enough material at these distances from the host star for these objects to have *formed* there. Although they will almost certainly not have formed at such large separations, they may have been scattered out to large radii through gravitational interactions (although this would result in large eccentricities), or migrated outwards through the disc, as was shown to be possible by Paardekooper & Mellema (2006) in non-isothermal discs.

2.5.2 Gravitational Instability Theory of Planet Formation

It is well known that stars form due to the gravitational instability in molecular clouds, and it was postulated by Kuiper (1951) that planets may also form through the same mechanism. More recently, this idea was revived by Boss (1997), who showed how protostellar discs may fragment to form gravitationally bound objects. In section 2.4.3.3, we discussed the Toomre criterion for gravitational instability

$$Q = \frac{c_{\rm s}\Omega}{\pi {\rm G}\Sigma}, \tag{2.171}$$

with numerical simulations showing that for non-axisymmetric instability, we require $Q < 1.5 - 1.7$ (Durisen et al., 2007). The Toomre parameter considers the stability of one patch of disc at any given time. Although $Q < 1.5 - 1.7$ is a

necessary condition for fragmentation, it provides us with no information about how the evolution of the system ought to proceed if it is to fragment into bound objects.

Considering, again, that equation 2.171 basically describes the balance between heating, cooling, self-gravity and rotation in the disc, we may ask at what point will the cooling dominate the heating, allowing the disc to collapse? To answer this, we begin by considering the dynamical timescale of a disc, given by $t_{\rm dyn} \sim \frac{1}{\Omega}$. This is the timescale of heating in the disc; spiral arms move through the disc, and regions that have cooled will collide with each other and reheat. If cooling is efficient, i.e. the cooling time is rapid, then cooled regions will gravitationally collapse before they have a chance to be reheated. This can be expressed as

$$t_{\rm cool} < \beta t_{\rm dyn}, \tag{2.172}$$

where β is a dimensionless parameter that captures the uncertainty in the efficiency of the cooling. This is more generally written as

$$\beta = t_{\rm cool}\Omega. \tag{2.173}$$

Helpfully, this cooling prescription can be incorporated analytically into the viscous prescription of accretion discs. We begin with the only relevant component of the viscous stress tensor, from the Navier-Stokes equation, for

circular motion[14]

$$T_{R\phi} = \nu \Sigma R \frac{d\Omega}{dR}, \tag{2.174}$$

then by using the Shakura & Sunyaev (1973) prescription for α-viscosity, where $\nu = \alpha c_{\rm s} H = \frac{\alpha c_{\rm s}^2}{\Omega}$, we have[15]

$$T_{R\phi} = \alpha c_{\rm s}^2 \Sigma \left| \frac{d \ln \Omega}{d \ln R} \right|. \tag{2.175}$$

In the disc, heating occurs as a result of viscous dissipation, and the dissipation rate, Q^+, per unit surface, is given by multiplying the viscous stress by the rate of shearing, $R\frac{d\Omega}{dR}$, such that

$$Q^+ = T_{R\phi} \left| \frac{R d\Omega}{dR} \right| = \alpha \Sigma c_{\rm s}^2 \Omega \left| \frac{d \ln \Omega}{d \ln R} \right|^2. \tag{2.176}$$

We can now define the radiative cooling , Q^-, per unit surface by using the internal energy per unit surface, U, such that

$$Q^- = \frac{U}{t_{\rm cool}} = \frac{\Sigma c_{\rm s}^2}{\gamma(\gamma - 1) t_{\rm cool}}, \tag{2.177}$$

[14]In general, shearing stress is given by $\mu \frac{dv_\phi}{dr}$, where μ is dynamic viscosity. Using $\mu = \nu\Sigma$, where ν is kinematic viscosity, we can express the shearing stress for our particular set of circumstances as $T_{R\phi} = \nu \Sigma R \frac{d\Omega}{dR}$

[15]Here we have made use of the chain rule, twice, so we have $\frac{d \ln y}{d \ln x} = \frac{d \ln y}{dx} \cdot \left(\frac{d \ln x}{dx} \right)^{-1} = \frac{d \ln y}{dx} \cdot \left(\frac{1}{x} \right)^{-1} = \frac{d \ln y}{dy} \cdot \frac{dy}{dx} \left(\frac{1}{x} \right)^{-1} = \frac{x}{y} \frac{dy}{dx}$.

where γ is the ratio of specific heats. Equating Q^+ and Q^-, we have

$$\alpha = \left|\frac{d\ln\Omega}{d\ln R}\right|^{-2} \frac{1}{\gamma(\gamma-1)t_{\rm cool}\Omega}. \tag{2.178}$$

Now, using[15]

$$\frac{d\ln\Omega}{d\ln R} = \frac{R}{\Omega}\frac{d\Omega}{dR}, \tag{2.179}$$

and calculating $\frac{d\Omega}{dR} = \frac{d}{dr}(\frac{GM}{R^3})^{\frac{1}{2}} = -\frac{3}{2}$, we finally arrive at (Gammie, 2001)

$$\alpha = \frac{4}{9}\frac{1}{\gamma(\gamma-1)t_{\rm cool}\Omega}. \tag{2.180}$$

The α that appears in equation 2.180 is the Shakura-Sunyaev α, and can therefore be thought of as an effective stress (characterising the actual stress tensor) that describes the efficiency of angular momentum transport. If α becomes too high, the disc cannot dissipate the increased stress rapidly enough to maintain a state of marginal instability, and so the disc fragments. Numerical simulations have shown this critical value of α to be ~ 0.06 (Gammie, 2001; Rice et al., 2005a). It is therefore not possible for regions of discs with very long cooling times to fragment, and this generally limits fragmentation due to the gravitational instability to the outer ~ 50 au and beyond (Rafikov, 2005).

However, there have been recent hydrodynamics simulations by Meru & Bate (2012) that show that fragmentation may occur for much smaller critical values of α than previously thought, with some previous work Meru & Bate (2011) suggesting that as resolution increases, this α value continues to decrease. If correct, this implies that fragmentation may occur in all discs, given sufficient

time to do so. If this is true, then our fundamental understanding of viscous accretion disc theory is incorrect, since our current understanding is that a disc must cool at a rapid enough rate, relative to its dynamical timescale, so that the cooling is not balanced by collisional heating. If fragmentation can occur in all discs given enough time to do so, then our fundamental understanding of the underlying balance between cooling and heating is wrong.

In the following chapter, I outline a particle method of modelling fluid dynamics, and discuss our investigation into this so-called "convergence problem" in section 3.10, and outline the likely solution, as suggested by Rice et al. (2012a), and shown by Rice et al. (2014a), in the conclusion to Chapter 3, found in section 3.11.

Programming today is a race between software engineers striving to build bigger and better idiot-proof programs, and the Universe trying to produce bigger and better idiots. So far, the Universe is winning.

Rick Cook, The Wizardry Compiled

3

Smoothed Particle Hydrodynamics Simulations of Fragmenting Protostellar Discs

3.1 Motivation

This chapter contains preliminary work undertaken into the issue of convergence of the fragmentation boundary in smoothed particle hydrodynamics (SPH) simulations of protostellar discs. With this in mind, I outline the theory behind

SPH, a Lagrangian method of modelling fluids. I begin with an introduction to SPH, followed by the discretisation of the equations of hydrodynamics. I discuss artificial viscosity, and artificial conductivity, before discussing the work undertaken into the lack of convergence of the fragmentation boundary of self-gravitating protostellar discs with increasing resolution. Ultimately, this particular line of enquiry did not avail us of a solution to the convergence problem (we discuss this in the conclusion to this chapter), however it did highlight that when faced with a numerical problem, the selected implementation of SPH is important. This work is preliminary, with potential for development and further investigation.

3.2 Introduction

Astrophysical flows are in general complex; there are few situations in which the equations of hydrodynamics have a direct analytical solution. However, most of the structures in the visible universe are formed in some way due to the interactions between gas dynamics, gravity and radiation. We therefore rely on numerical hydrodynamic simulations to understand these physical processes. Broadly speaking, there are two approaches to solving this problem. The first is *grid based*. Early work into this field relied upon fixed grid, finite-difference methods, splitting the physical domain of the simulation into grid cells, and calculating the flux across each boundary. However, fixed grids are not suitable for problems with a large dynamical range, such as the gravitational collapse of regions of a protostellar disc into bound objects. Computing this problem on a fixed grid leads to a large resolution difference across the simulation.

This led to the development of adaptive-mesh refinement (AMR) methods, whereby the grid is dynamically refined on some resolution criteria, and each grid level is contained entirely within another rectangular grid, except for the most coarse grid (Berger & Colella, 1989; Berger & Oliger, 1984). There are difficulties with AMR grids however, in that they are non Galilean invariant, which may lead to numerical mass diffusion across the boundary between cells. Additionally, AMR needs many levels to resolve large density contrasts, which becomes computationally expensive.

The second approach to solving numerical hydrodynamics problems is particle based, where the fluid particles directly discretise the fluid, and then the equations of hydrodynamics are solved directly between pairs of particles. This means that it is naturally conservative, and has a large dynamic density range. This method has ultimately become known as smoothed particle hydrodynamics (SPH), and was formulated by Lucy (1977) and Gingold & Monaghan (1977). Although it has advantages over AMR, it does come with its own set of problems. For example, it has limited adaptivity, only being able to adapt to density, which means poor resolution in regions of low density. AMR can be configured to minimize this error in low density regions. Shocks are not naturally captured in SPH, and require the use of artificial viscosity to prevent particle interpenetration in shock regions, and this leads to poor capturing of shocks. Additionally, it has recently been shown that true numerical convergence in SPH is more complicated than initially thought. It has generally been assumed that in the joint limit $N \rightarrow \inf$, $h \rightarrow 0$, where N is number of particles and h is the smoothing length scale, then numerical convergence will be achieved. However, Zhu et al. (2015) showed that

if the number of neighbours $N_{\rm nb}$ remains fixed, then convergence is not achieved as $N \to \inf$ and $h \to 0$, but actually requires $N \to \inf$, $h \to 0$ and $N_{\rm nb} \to \inf$.

However, we use SPH for its naturally conservative nature, and the relative ease with which it handles large density contrasts. We now outline the fundamental ideas behind SPH, and develop the Lagrangian equations of motion for fluid dynamics from their Eulerian forms that we outlined in section 2.2.1.

3.2.1 The Core Principle of SPH

Before we derive the Lagrangian equations of motion for SPH, we take pause for a moment and consider the most fundamental concept. Ultimately, given a distribution of point mass particles, we want to know how to obtain the density at any given point from this arbitrary distribution of particles. This is because the equations of motion in SPH, at least without any additional dissipative terms, can be derived entirely from the density. To see the full derivation of the discretised equations of hydrodynamics in this manner, we refer the reader to Price (2012), however to illustrate our point, we show that for the Lagrangian

$$L = T - V, \tag{3.1}$$

where T is kinetic energy and V is potential energy, then for a system of point masses the discrete version of this is simply

$$L = \sum_i m_i \left(\frac{1}{2} v_i^2 - u_i(\rho_i, s_i) \right), \tag{3.2}$$

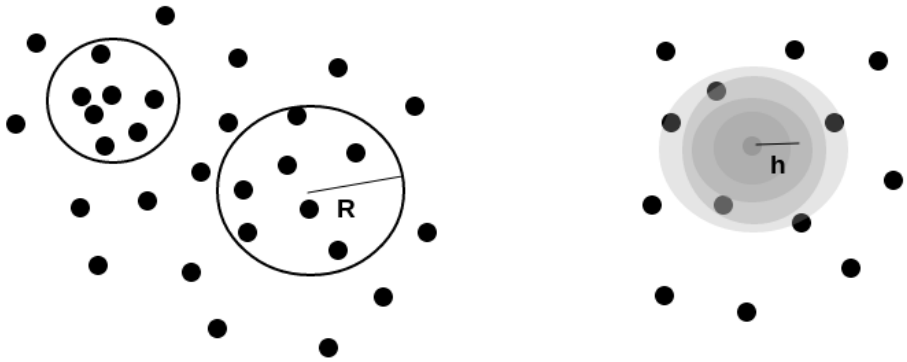


Figure 3.1: Density determination using point particles. Left: adapting the volume to keep a constant number of neighbours. All neighbours weighted evenly. Right: Density computed via a weighted sum, decreasing the weight as we step outwards in radius from the central particle.

where m_i is particle mass, v_i is particle velocity, and u_i is internal energy of the particle, which is a function of density, ρ_i, and entropy, s_i. In continuous form, this is

$$L = \int [\rho v^2 - \rho u(\rho, s)] \, dV, \tag{3.3}$$

and we can then see the dependence of L upon density.

Now that we have outlined the importance of this property, we can show how it is determined. The left hand side of Figure 3.1 shows how the density can be determined from a group of particles, using an adaptive volume size so that the number of neighbour particles, N_{nb}, remains the same. This naturally adjusts the resolution in sparse or dense regions. The density is then simply computed by

$$\rho = \frac{\sum_i^{N_{\text{nb}}} m_i}{V}, \tag{3.4}$$

where m_i is the particle mass, and V is the volume under consideration. However,

this creates a noisy density estimate, since particles will often sit on the boundary of being included in the volume sphere. It therefore makes sense to weight this estimate such that particles near the edge of this volume, known as the neighbour sphere, contribute less to the density estimate than those closest to the centre. The density would then be given by

$$\rho(\mathbf{r}) = \sum_{i}^{N_{\rm nb}} m_i W(\mathbf{r} - \mathbf{r}_i, h), \tag{3.5}$$

where W is the weight function, known as the *smoothing kernel*, and h is the *smoothing length*, essentially a scale parameter which dictates the rate of fall off of function W. Since it makes sense to equally resolve both dense and sparse regions, the smoothing length is set by simultaneously solving (Price, 2008; Price & Monaghan, 2004)

$$\rho(\mathbf{r}_a) = \sum_{b} m_b W(\mathbf{r}_a - \mathbf{r}_b, h_a); \quad h(\mathbf{r}_a) = \eta \left(\frac{m_a}{\rho_a}\right)^{\frac{1}{D}}, \tag{3.6}$$

where D is the number of dimensions in the simulation and η parameterises the smoothing length in units of mean particle spacing, $(m/\rho)^{1/D}$. This approach, to date, has the best accuracy (Price, 2008), but does involve the addition of extra terms to the discretised SPH equations in order to be fully conservative. This is known as grad-h SPH, and we discuss this further in section 3.10.3.3.

However, there are many ways of adaptively refining the smoothing length. One such approach, as already discussed, is to impose that the number of neighbours in the neighbour sphere around a particle must remain constant, although this has the disadvantage of greatly decreasing the smoothing length of a particle

approaching a dense region. In most "standard" implementations of SPH (see, e.g., Monaghan 1992), the smoothing length of each particle, i, is evolved through

$$\frac{dh_i}{dt} = -\frac{h_i}{D\rho_i}\frac{d\rho_i}{dt}. \tag{3.7}$$

In this case, a symmetric kernel must be used when obtaining any interpolated quantity, which can be obtained simply by using the arithmetic mean of the smoothing length. For example, interpolating for the density at location b, using the sum over all particles a, requires the average smoothing length, $h_{ab} = \frac{1}{2}(h_a + h_b)$.

A Gaussian is the most natural choice for W, since it decreases monotonically with distance and is symmetrical. However, in reality a Gaussian is not practical, since by definition at no point is it zero, therefore requiring its computation at every point in the domain to get an accurate density estimate.

A more practical choice, therefore, is a spline function, although there are many Kernel functions that also fit this. Since this spline needs to be continuous in its first and second derivatives, a third order B-spline is commonly found in SPH, an example in 3D is given by

$$W(q = \frac{\mathbf{r} - \mathbf{r}'}{h}) = \frac{1}{h^3\pi}\begin{cases} \frac{1}{4}(2-q)^3 - (1-q)^3 & 0 \leq q < \frac{1}{2} \\ \frac{1}{4}(2-q)^3 & 1 \leq q < 2 \\ 0 & q \geq 2. \end{cases} \tag{3.8}$$

In fact, we have now arrived at a crucial point. At the heart of SPH, any function

$f(\mathbf{r})$ can be approximated by using a kernel, such that the approximation to the function, $\tilde{f}_h(\mathbf{r})$, is given by

$$\tilde{f}_h(\mathbf{r}) = \int f(\mathbf{r'})W(\mathbf{r}-\mathbf{r'},h)\ dV. \tag{3.9}$$

As mentioned, the Kernel can take a variety of forms, but to make physical sense should fulfil

$$\lim_{h\to 0} \tilde{f}_h(\mathbf{r}) \to f(\mathbf{r'}) \tag{3.10}$$

and

$$\int W(\mathbf{r}-\mathbf{r'},h)\ dV = 1. \tag{3.11}$$

We are now in a position to begin to derive the discrete equations of SPH.

3.3 Discretising the Hydrodynamics Equations

Having derived the equations of hydrodynamics earlier in Eulerian form (see section 2.2.1, page 14), we now simply state their Lagrangian equivalent, making use of the relationship between the Lagrangian time derivative $^{\mathrm{d}}/_{\mathrm{dt}}$, and the Eulerian time derivative $^{\partial}/_{\partial t}$, given by (Rosswog, 2009)

$$\frac{d}{dt} = \frac{dx^i}{dt}\frac{\partial}{\partial x^i} + \frac{\partial}{\partial t} = \mathbf{v}\cdot\nabla + \frac{\partial}{\partial t}, \tag{3.12}$$

gives in Lagrangian form the *continuity equation*

$$\frac{d\rho}{dt} = -\rho\nabla\cdot\mathbf{v}, \tag{3.13}$$

the *momentum equation*

$$\frac{d\mathbf{v}}{dt} = -\frac{\nabla P}{\rho} + \mathbf{F}_b, \tag{3.14}$$

and finally the energy equation

$$\frac{du}{dt} = \frac{P}{\rho^2}\frac{d\rho}{dt} = -\frac{P}{\rho}\nabla \cdot \mathbf{v}. \tag{3.15}$$

3.3.1 The Continuity Equation

We begin by shortening our notation, for brevity, such that any vector $\mathbf{r}_{ab}$ is the relative vector between an SPH particle a, and its neighbours, b. Since the density in SPH is estimated from equation 3.5, then the continuity equation can be discretised as

$$\frac{d\rho_a}{dt} = \sum_b m_b \frac{dW_{ab}}{dt}, \tag{3.16}$$

where

$$W_{ab} = \frac{1}{2}[W(\mathbf{r}_a - \mathbf{r_b}, h_a) + W(\mathbf{r}_a - \mathbf{r_b}, h_b)]. \tag{3.17}$$

However, we can make use of the time derivative of the kernel, given by

$$\begin{aligned}
\frac{dW}{dt} &= \frac{\partial W_{ab}}{\partial r_{ab}}\frac{dr_{ab}}{dt} \\
&= \frac{\partial W_{ab}}{\partial r_{ab}}\frac{(\mathbf{r}_a - \mathbf{r}_b)\cdot(\mathbf{v}_a - \mathbf{v}_b)}{r_{ab}} \\
&= \frac{\partial W_{ab}}{\partial r_{ab}}\hat{\mathbf{e}}_{ab}\cdot\mathbf{v}_{ab} \\
&= \mathbf{v}_{ab}\cdot\nabla_a W_{ab}
\end{aligned} \tag{3.18}$$

such that

$$\frac{d\rho_a}{dt} = \sum_b m_b \mathbf{v}_{ab} \cdot \nabla_a W_{ab}, \tag{3.19}$$

where ∇_a is the gradient taken at a.

3.3.2 The Momentum Equation

The momentum equation, when discretised directly, does not, strictly speaking, conserve Newton's third law. Since

$$\frac{d\mathbf{v}_a}{dt} = -\frac{1}{\rho_a} \sum_b \frac{m_b}{\rho_b} P_b \nabla_a W_{ab}, \tag{3.20}$$

we can see that if we compute the force using the derivative of this equation, we have a dependence upon P. Since, in general, P_a does not equal P_b, this equation does not conserve total momentum, because $m\frac{d\mathbf{v}_a}{dt}$ will not equal $m\frac{d\mathbf{v}_b}{dt}$. Instead, we use the slightly more complicated version (by making use of the product rule, c.f. Rosswog 2009),

$$\frac{d\mathbf{v}_a}{dt} = -\sum_b m_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2} \right) \nabla_a W_{ab}, \tag{3.21}$$

since this ensures the pressure term is now symmetric in a and b.

3.3.3 The Energy Equation

We construct directly from equation 3.15 the discretised form of the energy equation,

$$\frac{du_a}{dt} = \frac{P_a}{\rho_a^2}\frac{d}{dt}\left(\sum_b m_b W_{ab}\right) \tag{3.22}$$

$$= \frac{P_a}{\rho_a^2}\sum_b m_b \mathbf{v}_{ab}\cdot\nabla_a W_{ab}, \tag{3.23}$$

where we have again made use of the time derivative of the kernel given in equation 3.18.

3.4 Artificial viscosity

When simulating astrophysical fluids, regions of large pressure or density difference, known as shocks, are almost unavoidable. On very small length scales, i.e. comparable to the mean free path of the gas, these shock regions are continuous due to the *physical viscosity* present. However, simulations of these events take place on much larger length scales, meaning that the values of these macroscopic variables appear to be discontinuous. This is shown in Figure 3.2. The left-hand panel shows a shock region that is resolvable on simulation length scale, with the shock spread over length Δx. The right-hand panel shows the same shock, but the smallest resolution of the simulation is larger than Δx, so the shock appears as a discontinuity. This discontinuity results in unphysical oscillations behind the shock front, shown in Figure 3.3. Both panels show a 1D Sod shock tube problem (Sod, 1978), both implemented with "standard" SPH

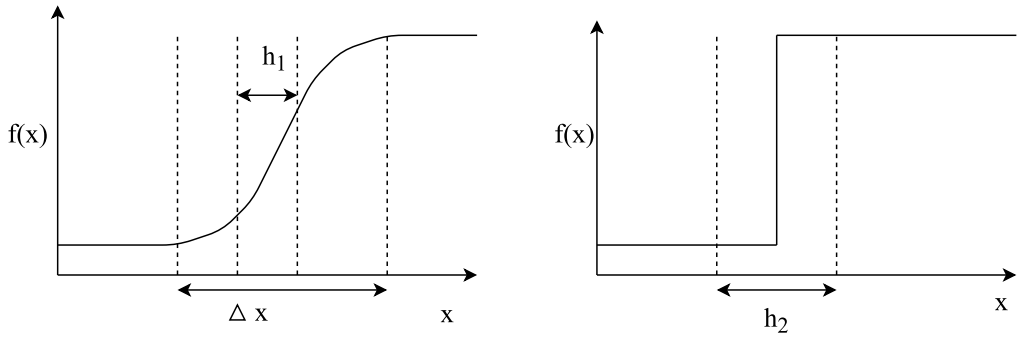


Figure 3.2: Left shows a shock region that is resolvable on simulation length scale, since the change in $f(x)$ happens in width Δx, which is larger than the smallest resolvable scale, h_1 in the simulation. Right shows the same shock region, but with the smallest resolvable scale in the simulation, h_2, larger than Δx. This shock now appears as a discontinuous function.

artificial viscosity (see, e.g.,Monaghan 1992). We can see that on the left, there is unphysical oscillation in the pressure. Generally speaking, there are three solutions to this problem:

1. Increase the resolution in that region, by reducing the smoothing length h.
2. Smooth the shock over a region larger than h, making the shock resolvable.
3. Use the exact solution to the Riemann problem.

Item 1 is generally undesirable, since it can introduce additional noise in an already messy region. Item 3 removes the problem altogether, by treating two interacting particles as left and right states of the Riemann problem (see, for example, Cha & Whitworth 2003; Inutsuka 1994, 2002; Murante et al. 2011). However, Riemann solvers are computationally expensive, since the Riemann problem needs to be solved for every pair of particles near the shock boundary. Additionally, it is difficult to do if a more complicated equation of state is used

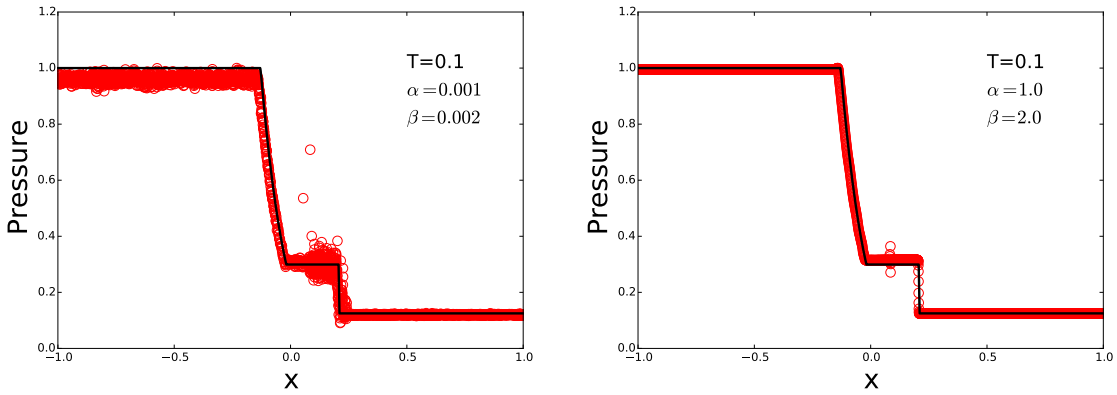


Figure 3.3: Both panels show the 1D Sod shock tube problem, evolved to $T = 0.1$, with the analytical solution plotted in solid black line. Open red circles show the SPH solution. Both plots show pressure as a function of distance, and were initiallised with left state $\rho_{\rm L} = 1.0$, $p_{\rm L} = 1.0$, $v_{\rm L} = 0.0$ and right state $\rho_{\rm R} = 0.1$, $p_{\rm R} = 0.125$, $v_{\rm R} = 0.0$. The left-hand side shows the simulation implemented with a small amount of artificial viscosity, which allows unphysical oscillations in the shock region. The right-hand panel shows the same problem but with a higher amount of artificial viscosity, and aside from the pressure "blip" at $x \sim 0.1$, we have a good agreement with the analytical solution.

(Price, 2005). For this reason, we select item 2 as our chosen solution, and proceed.

To smooth the shock over a resolvable region, we add a small amount of *artificial viscosity* to the simulation, which acts to create entropy in the shock region, much as is done by physical viscosity, but on a numerically resolvable scale (Rosswog, 2009). We add this viscosity by adding pressure-like terms to the fluid equations, but so as not to introduce numerical effects it should always be dissipative (transferring kinetic energy into internal energy), it should also be absent in shockless conditions, such as rigid body rotation, and should also disappear in uniform compression. This motivated Von Neumann & Richtmyer (1950) to introduce viscosity that can spread the shock to resolvable scales, and

is now commonly used in SPH. The momentum equation is now modified to (Monaghan, 1992)

$$\frac{d\mathbf{v}_a}{dt} = -\sum_b m_b \left(\frac{P_a}{\rho_a^2} + \frac{P_b}{\rho_b^2} + \Pi_{ab} \right) \nabla_a W_{ab}, \tag{3.24}$$

where

$$\Pi_{ab} = \begin{cases} -\frac{\alpha \bar{c}_{ab} \mu_{ab} + \beta \mu_{ab}^2}{\bar{\rho}_{ab}} & \mathbf{v}_{ab} \cdot \mathbf{r}_{ab} \leq 0 \\ 0 & \mathbf{v}_{ab} \cdot \mathbf{r}_{ab} > 0. \end{cases} \tag{3.25}$$

Here, $\bar{\rho}_{ab}$ is the average density, $\bar{c}_{ab}$ is the average sound speed, α and β have an acceptable range of values from the literature, with numerical experiments suggesting $\alpha \sim 1$ and $\beta \sim 2$ (and always $\beta = 2\alpha$), and μ_{ab} given by

$$\mu_{ab} = \frac{h \mathbf{v}_{ab} \cdot \mathbf{r}_{ab}}{r_{ab}^2 + \epsilon \bar{h}_{ab}^2}. \tag{3.26}$$

The quadratic in velocity β term is dominant in the case of large velocity differences (i.e. $\sim$ Mach number shocks), preventing particle interpenetration by producing an artificial pressure. Particle interpenetration is undesirable because, if it occurs, the fluid at a particular location in space may be described by multiple values. For example, imagine two particles travelling with velocities such that they will collide at the location in space (x, y, z). Particle a is travelling with velocity $\mathbf{v}_a$ and particle b is travelling with velocity $\mathbf{v}_b$. If particle interpenetration occurs, the flow velocity of the fluid at (x, y, z) will be described by two values, $\mathbf{v}_a$ and $\mathbf{v}_b$, since both particles occupy the same location in space. This is obviously unphysical, and is prevented by the addition of artificial viscosity. The linear in velocity α term is dominant in small velocity differences. Similarly, the energy

equation is also modified, such that

$$\frac{du_a}{dt} = \frac{P_a}{\rho_a^2} \sum_b m_b \mathbf{v}_{ab} \cdot \nabla_a W_{ab} + \frac{1}{2} \sum_b m_b \Pi_{ab} \mathbf{v}_{ab} \cdot \nabla_a W_{ab}. \tag{3.27}$$

The result of this is shown in the right-hand panel of Figure 3.3, where we can see that the SPH values, shown in red cirles, are now much closer to the analytical solution, shown as the solid black line.

3.5 Gravity in SPH

3.5.1 Avoiding Singularities in Force Calculations

When modelling fluid dynamics, in almost all situations other than the astrophysical, gravity will only be included in the simulation as an external force. However, self-gravity is extremely important in astrophysical flows, and is included accordingly. In N-body simulations, the force on particle i due to gravity is given by

$$F_i = \sum_{j=1}^{N} \frac{\mathrm{G} m_i m_j}{|\mathbf{r}_i - \mathbf{r}_j|^2 + \epsilon^2}, \tag{3.28}$$

where ϵ is the *softening length*, which prevents the force from becoming infinite at zero particle separation. However, SPH can make use of the smoothing kernel to prevent this infinite force. As before, we consider the mass of particle i as being smeared out in its local smoothing volume, to give a density profile such that

$$\rho(\mathbf{r}) = m_i W(|\mathbf{r} - \mathbf{r}_i|, h_i) \tag{3.29}$$

where W is the smoothing kernel. If we now consider particle j, which is inside the smoothing volume of particle i, then we have

$$m_i(< r_j) = 4\pi \int_0^{r_j} r^2 \rho(r)\, dr = 4\pi m_i \int_0^{r_j} r^2 W(|\mathbf{r} - \mathbf{r}_i|, h_i)\, dr \tag{3.30}$$

to describe the mass of particle i that is between $\mathbf{r}_i$ and $\mathbf{r}_j$. If we now replace m_i with $m_i(< \mathbf{r}_j)$, then our equation for gravitational force can be written as

$$F_i = \sum_{j=1}^{N} \frac{\mathrm{G} m_i(< \mathbf{r}_j) m_j}{|\mathbf{r}_i - \mathbf{r}_j|^2}, \tag{3.31}$$

then at $\mathbf{r}_i = \mathbf{r}_j$, $F_i = 0$, i.e. the force vanishes, removing the singularity at $|\mathbf{r}_i - \mathbf{r}_j| = 0$. Since we have removed the singularity, there is no need for a softening length. If the particles do not overlap, then $m_i(< r_j) = m_i$. This process is known as *kernel softening*.

3.5.2 The Use of Tree Structures

Calculating gravitational forces is computationally expensive, $\mathcal{O}(N)^2$, if done using a particle-particle approach. Prior to 1986, in order to reduce computational costs to $\mathcal{O}(N \log N)$, it was common practice to fit a model with free parameters to the global gravitational potential, and subsequently propagate all particles in this potential before repeating the procedure. However, this method involved implicit assumptions about the geometry of the system and would not model local interactions correctly.

When Barnes & Hut (1986) published their seminal paper detailing a new force

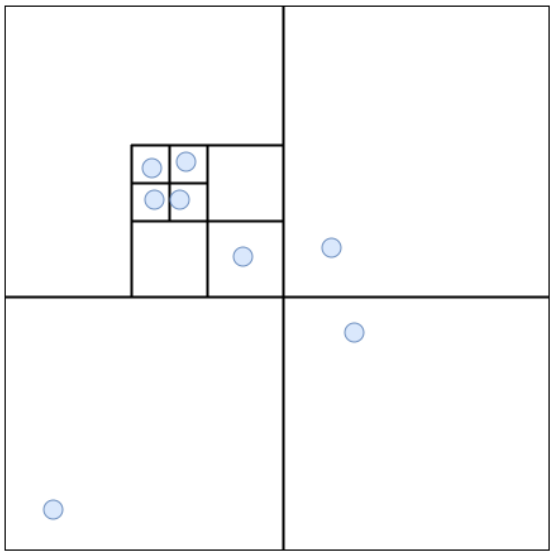


Figure 3.4: This is a quadtree, the 2D equivalent of an octree. The tree cells subdivide into four children, until there are no cells containing more than one particle.

calculation algorithm, it is not an overstatement to say that it revolutionised the subsequent development of N-body codes, and indeed any code that must calculate the effects of self-gravity. The basic idea is to calculate the long range gravitational forces by aggregating all of the effects of a group of particles, a large distance away, into one particle. This principle is analogous to calculating the gravitational force between two galaxies; we do not need to understand the internal distribution of stars, *if* we are sufficiently far away. In practice, this is done by splitting the computational domain into an octree (in 3D) or a quadtree (in 2D), which is shown in Figure 3.4. In an octree, we begin with a cube that bounds the entire simulation domain. The cube is subsequently divided into 8 cells, and these are known as child cells. If these cubes have more than one particle in, they are further subdivided into eight grandchild cells, and the process is repeated until there is only one particle left in each cell, at which point, these cells are known as *leaf cells*. In reality, it is likely that more than one particle

will be allowed to remain in a leaf cell, the criterion will be selected based upon tests that show sufficient adherence to accuracy without compromising speed.

To calculate the force on particle i, we begin at the root cell, and walk down the tree until we reach the leaf cell that contains particle i. Each cell is of size l, and D is the distance between the cell's center of mass and particle i. At each cell, we check if $l \lesssim D$ (in reality, we check if $l \leq \theta D$, where $\theta \sim 1$ sets the tolerance). If l *is* less than D, then we add the force contribution from the particles in that cell as though they are all located at the center of mass of the cell. If, however, $l \gtrsim D$, we continue to descend down the tree. If we hit a leaf cell, for which $l \gtrsim D$, then individual particle-particle interactions are calculated for all the particles in the leaf.

3.5.3 Sink Particles

SPH simulations are often used to simulate the collapse and subsequent formation of dense objects, such as brown dwarfs and gas giant planets in simulations of protostellar disc fragmentation. However, particles in very dense regions will have very small timesteps, which essentially means that the simulation will grind to a halt once high density regions have been created.

Instead, these particles can be replaced by one pointmass particle, usually known as a *sink* particle (Bate et al., 1995). These sink particles do not interact through any force but gravity, and these sinks also accrete any other particle that comes inside the radius of accretion, $r_{\rm acc}$, so long as the particle is gravitationally bound to the sink, and so long as the specific angular momentum of the particle,

calculated relative to the sink, is less than the angular momentum required for the particle to form a circular orbit around the sink at a radius of $r_{\rm acc}$. This ensures that no particles are accreted that would be likely to leave the sink.

Once the particle has been accreted, its mass and linear momentum are added to the sink particle, and then the position of the sink particle is modified, so that it now occupies the center of mass of the previous sink-particle system. Finally, the spin of the sink is modified to account for the angular momentum of the particle it has just accreted.

Sink particles may be included at the start of the simulation, for example, as a central star in a protostellar disc. But they are particularly helpful when they can be created dynamically. To do so, we typically ensure that several conditions are met. First, the particle must be more dense than some threshold value, say, $\sim 10^5$ times the initial density. This triggers the sink creation process. The tests are as follows:

1. Check that ~ 50 neighbours exist within one accretion radius.
2. $\alpha \leq \frac{1}{2}$, where α is the ratio of thermal energy to the magnitude of gravitational energy.
3. $\alpha + \beta \leq 1$, where β is the ratio of rotational energy to the magnitude of gravitational energy.
4. The total energy of the particles must be negative, which ensures that the sink is gravitationally bound.
5. The divergences of the accelerations of the particles are less than zero, which

ensures that the particles are in the act of collapsing. Otherwise, they may be in the process of being tidally disrupted, and the sink could shear away.

3.6 Timestepping

When simulating a hydrodynamical system, to ensure that the fluid behaves correctly, the equations of hydrodynamics must be integrated on a timescale, known as the *timestep*, that is shorter than the shortest timescale of interest of the fluid. In the absence of an external force, this is known as the *Courant Friedrichs Lewy* (CFL) condition (Courant et al., 1928); enforcing it prevents the numerical speed of information propagation exceeding the physical speed. This condition is described by (Press et al., 1992)

$$\delta t < \frac{\delta x}{c_{\rm s}}, \tag{3.32}$$

where δx is the minimum length scale that may be resolved, and $c_{\rm s}$ is the sound speed of the medium. In practice, the timestep, δt, is chosen based on a combination of the CFL condition, the force terms, and the viscous diffusion terms, given by (Monaghan, 1989; Monaghan, 1992)

$$\delta t = k \ \min(\delta t_{\rm f} + \delta t_{\rm cv}), \tag{3.33}$$

where the force condition is calculated from the local force per unit mass, $\mathbf{f}_{\rm a}$, and local smoothing length, $h_{\rm a}$,

$$\delta t_{\rm f} = \min_{\rm a} \left(\frac{h_{\rm a}}{|\mathbf{f}_{\rm a}|} \right), \tag{3.34}$$

and the combined Courant and viscous condition is

$$\delta t_{\rm cv} = \min_{\rm a} \frac{h}{c_{\rm a} + 0.6(\alpha c_{\rm a} + \beta \max_{\rm b} \mu_{\rm ab})}, \tag{3.35}$$

where α and β are the standard artificial viscosity parameters described in section 3.4, and $\mu_{\rm ab}$ is defined in equation 3.26. Although it is easier to integrate all particles using the same timestep, in reality this is rarely the most efficient method when modelling astrophysical situations, since there may be regions of the simulation that evolve much more rapidly than others. In this case, it makes sense to allow particles to have their own timesteps, and synchronise them with the rest of the simulation at regular intervals.

For example, allowing each particle to have a timestep t_i, with an initial timestep, $t_{\rm max}$, for all particles in the simulation, will give

$$t_i = \frac{t_{\rm max}}{2^{n_i}}, \tag{3.36}$$

where $n_i \geq 0$ is the *time bin*. If $n_i = 0$, then each $t_i = t_{\rm max}$, i.e. each particle is in sync with the entire simulation. If $n_i = 1$, then particle i will synchronise with the main simulation's timestep, $t_{\rm max}$, every $2t_i$ timesteps.

3.6.1 Time Integration

Consultation of any numerical recipes book will reveal to the reader a wide variety of algorithms for solving differential equations. We give two common examples, the Runge-Kutta, and Leapfrog methods. In a Runge-Kutta approach, a solution

is propagated over an interval by using multiple Euler steps, given by

$$y_{n+1} = y_n + hf(x_n, y_n), \tag{3.37}$$

where h is the stepsize, and $x_{n+1} = x_n + h$. Euler's method is *first order accurate*, so-called because its error term is proportional to the step-size squared, i.e. $\mathcal{O}h^2$. By convention, a method is generally called nth order accurate if its error terms are $\mathcal{O}(h^{n+1})$. However, by taking an additional midpoint step, the accuracy can be increased to second order, i.e.

$$y_{n+1} = y_n + hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}hf(x_n, y_n)) + \mathcal{O}(h^3). \tag{3.38}$$

In fact, this can be progressively continued to obtain smaller errors. One of the most common found in SPH integration schemes is fourth order Runge-Kutta. However, the Runge-Kutta method in general does not conserve energy, allowing the system to drift away from the true solution over time. This is, as one can imagine, entirely undesirable when integrating orbital dynamics[16].

However, the Leapfrog integrator does conserve energy, and so it is particularly useful in simulations that are integrated for long periods of time. It is so-called because the position and velocity calculations leap over one another. To see this,

[16]The reason for this is complex, but the basic idea is that Runge-Kutta algorithms are not symplectic, which means area is not conserved in phase-space (position-momentum), between time t_i and time t_{i+1} . However, integrating Newton's equations by a finite amount of time *is* symplectic, hence the drift away from the real solution.

consider first a simulation in only 1D. The equations of motion are then

$$\frac{dy}{dt} = v \tag{3.39}$$

and

$$\frac{dv}{dt} = F(y), \tag{3.40}$$

where F is the force on the particle when it is at location y. For this, the Euler method give

$$y_1 = y_0 + hv_0. \tag{3.41}$$

However, we know that a derivative calculated at a point poorly approximates the derivative of a line joining two points, and a better approximation is given by

$$y_1 = y_0 + hv_{\frac{1}{2}}. \tag{3.42}$$

This can then be used to step the velocity forwards in time, so we have

$$v_{\frac{3}{2}} = v_{\frac{1}{2}} + hF(x_1), \tag{3.43}$$

then we can step forward in y with

$$y_2 = y_1 + hv_{3/2}, \tag{3.44}$$

and so on. We therefore have, for leapfrog integration,

$$y_{n+1} = y_n + hv_{n+\frac{1}{2}} v_{n+\frac{3}{2}} = v_{n+\frac{1}{2}} + hF(y_{n+1}). \tag{3.45}$$

Leapfrog integrators, while being no more accurate than Runge-Kutta integrators (this particular example is second-order accurate), conserve global quantities, like energy. Strictly speaking, we have a slight problem, since we need an expression for $v_{n+\frac{1}{2}}$ in order to begin. We can approximate this with a half-step Euler, while not decreasing the overall order of the method since we only do this once. We now have

$$v_{n+\frac{1}{2}} = v_n + \frac{1}{2}hF(y_n).y_{n+1} = y_n + hv_{n+\frac{1}{2}}v_{n+1} = v_{n+\frac{1}{2}} + \frac{1}{2}hF(y_{n+1}), \qquad (3.46)$$

where we have split equation 3.45 into equal half steps.

3.7 Radiative Transfer in SPH

A full description of polychromatic, three-dimensional radiative transfer is currently beyond our computational abilities. As a result, we must make approximations that can capture individual features of radiative transfer, without the computational expense. One such example of this is the cooling formalism

$$\dot{u}_i = -\frac{u_i}{t_{\rm cool}}, \qquad (3.47)$$

where u_i is the internal energy of the particle, and $t_{\rm cool}$ is the cooling timescale, which dictates how rapidly the disc cools, given by

$$t_{\rm cool} = \frac{\beta}{\Omega}, \qquad (3.48)$$

where β is an integer, lower β indicates more rapid cooling. This prescription has been widely used in the investigation of gravitational instability (GI) in protostellar discs (see, eg, Gammie 2001 for the original analytic derivation, and Rice et al. 2003a for its use in SPH simulations), since its simplicity allows the gravitational instabilities to be characterised effectively. Despite its widespread adoption, this cooling parameterisation only captures energy loss from the disc.

In reality, energy loss from one particle, especially one embedded in the midplane of the disc, is likely to be transported, at least partially, to neighbouring particles. Since gravitational instabilities develop as the gas cools, this redistribution of energy is expected to alter the subsequent gas dynamics, and therefore the morphology of forming spiral arms. Consideration of these effects is, therefore, necessary, if we wish to accurately characterise the nature of GI in protostellar discs.

There are many methods for capturing the effects of radiative transfer in SPH (Mayer et al., 2007a; Stamatellos & Whitworth, 2005a; Whitehouse & Bate, 2004). However, we may ask why we do not directly incorporate the MCRT method discussed in section 2.2.4.1, for example, by following the method of Oxley & Woolfson (2003), whereby a full radiative transfer calculation is run in between hydrodynamical timesteps, and the resulting temperatures are applied to SPH particles for the next hydrodynamical timestep. However, this is computationally expensive, so much so that it, in practice, it is not feasible.

The method used in this thesis is not computationally (very) expensive, since it makes use of approximations to cooling and energy transfer, in what is known

as the hybrid radiative transfer method (Forgan et al., 2009). The basic idea is to model energy loss from the system using a more realistic prescription for the cooling time of a particle, i.e., $t_{\rm cool} = f(T, \tau)$, where T is temperature and τ is optical depth, at the same time as modelling the energy transport between particles.

In practice, this is achieved by using two separate methods at the same time, the polytropic cooling approximation of Stamatellos et al. (2007) to handle energy loss, and the flux limited diffusion approximation (Bodenheimer et al., 1990; Levermore & Pomraning, 1981; Whitehouse & Bate, 2004; Whitehouse et al., 2005), which models the transfer of energy between particles.

In the following sections, we first give a description of the polytropic cooling approximation, then the flux-limited diffusion approximation, before finally describing the hybrid method implemented by Forgan et al. (2009).

3.7.1 The Polytropic Cooling Approximation

We now describe the polytropic cooling formalism of Stamatellos et al. (2007). The polytropic approximation uses the particles's density, ρ_i, temperature, T_i and gravitational potential, ψ_i, to estimate a mean optical depth, $\bar{\tau}_i$, for the particle. This optical depth then regulates the radiative heating and cooling of the particle, i.e., it models the extent to which the particle is shielded from radiation, and how well the radiation from the particle will be trapped.

3.7.1.1 A Quick Primer on Polytropes and the Lane-Embden Equation

The interior of stars may be modelled as polytropic fluids of the form

$$P = K\rho^{\frac{n+1}{n}}, \tag{3.49}$$

where P is the pressure, K is a constant of proportionality, ρ is the density, and n is an integer. A polytrope is a solution to the Lane-Embden equation,

$$\frac{1}{\xi^2}\frac{d}{d\xi}\left(\xi^2\frac{d\theta}{d\xi}\right) + \theta^n = 0, \tag{3.50}$$

where ξ is the dimensionless radius, $R = \xi R_0$, and θ relates the density of the polytrope to the polytrope's central density, ρ_c, by

$$\rho = \rho_c\theta^n, \tag{3.51}$$

and therefore θ is also related to the pressure of the polytropic fluid by 3.49.

3.7.1.2 Polytropic Cooling

We begin with the assumption that the particle is embedded in a spherically symmetric polytropic pseudocloud, such that the properties of the cloud may be determined analytically through the Lane Emden equation. If we consider a particle embedded in the pseudocloud, the properties of the particle can be recovered through appropriate selection of central values and scale length, such

that

$$\rho_i = \rho_c \theta^n(\xi) \tag{3.52}$$

$$T_i = T_c \theta(\xi) \tag{3.53}$$

$$\psi_i = -4\pi \mathrm{G} \rho_c R_0^2 \phi(\xi), \tag{3.54}$$

where θ is the solution to the Lane-Embden equation for a polytrope of index n, and

$$\phi(\xi) = -\xi_B \frac{d\theta}{d\xi}(\xi_B), \tag{3.55}$$

for ξ_B describing the boundary of the polytrope. It is important to note that the position of the particle within the pseudocloud is not determined, but rather, a mass weighted average between all of the positions is taken. This is shown in Figure 3.5, where the SPH particle may be located anywhere within the pseudocloud.

Fixing n, and picking an arbitrary location inside the pseudocloud for the particle (i.e. $\xi < \xi_B$), then we have

$$\rho_c = \rho_i \theta^{-n}(\xi) \tag{3.56}$$

$$R_0 = \left(\frac{-\psi_i \theta^n(\xi)}{4\pi \mathrm{G} \rho_i \phi(\xi)} \right)^{\frac{1}{2}}. \tag{3.57}$$

Similarly, the central temperature of the cloud is then chose so that it reproduces

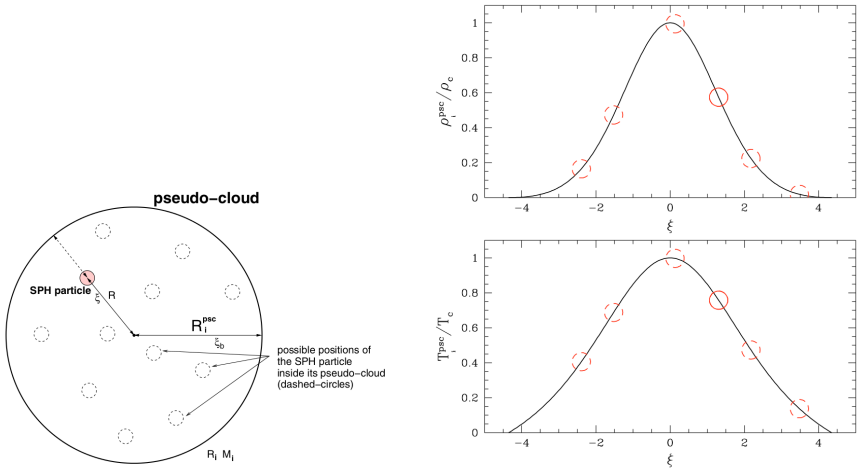


Figure 3.5: Figures taken from Stamatellos et al. (2007). Left is their Figure 1, right is their Figure 2. Left shows the schematic representation of the pseudo-cloud around a SPH particle. The location of the SPH particle inside its pseudo-cloud is not specified. Right shows density and temperature profiles for a polytropic pseudo-cloud with n = 2. The SPH particle could be located anywhere in the cloud (solid and dashed line circles).

the temperature of the particle at radius $R = \xi R_0$, and we have

$$T_c\theta(\xi) = T_i \tag{3.58}$$

$$T_c = T_i\theta^{-1}(\xi). \tag{3.59}$$

We can then calculate a column density from any given dimensionless radius to the boundary of the cloud, through

$$\begin{aligned}\Sigma_i(\xi) &= \int_{\xi'=\xi}^{\xi'=\xi_B} \rho_c\theta^n(\xi')R_0 d\xi' \\ &= \left(\frac{-\psi\rho_i}{4\pi \mathrm{G}\phi(\xi)\theta^n(\xi)}\right)^{\frac{1}{2}} \int_{\xi'=\xi}^{\xi'=\xi_B} \theta^n(\xi')d\xi',\end{aligned} \tag{3.60}$$

however, since ξ is unknown, we use instead a mass weighted average over all

possible values of ξ, out to the polytrope's boundary at $\xi = \xi_B$. This is given by

$$\bar{\Sigma}_i = \left(-\xi_B^2 \frac{d\theta}{d\xi}(\xi_B) \right)^{-1} \int_{\xi'=0}^{\xi'=\xi_B} \Sigma_i(\xi')\theta^n(\xi')\xi^2 \, d\xi'$$
$$\bar{\Sigma}_i = \zeta_n \left(\frac{-\psi_i \rho_i}{4\pi \mathrm{G}} \right)^{-1} \tag{3.61}$$

where $\left(-\xi_B^2 \frac{d\theta}{d\xi}(\xi_B) \right)^{-1}$ is the total dimensionless mass of the polytrope, and $\theta^n(\xi)\xi^2 \, d\xi$ is the dimensionless mass element between the dimensionless radius ξ and $\xi + d\xi$. Note that we have folded the integral into ζ_n, which only depends on the polytropic index n, and is given by

$$\zeta_n = \left(-\zeta_B^2 \frac{d\theta}{d\xi}(\xi_B) \right)^{-1} \int_{\xi=0}^{\xi=\xi_B} \int_{\xi'=\xi}^{\xi'=\xi_B} \theta^n(\xi') d\xi' \left(\frac{\theta^n(\xi)}{\phi(\xi)} \right)^{\frac{1}{2}} \xi^2 \, d\xi. \tag{3.62}$$

Stamatellos et al. (2007) showed that ζ_n is relatively insensitive to n, since $\zeta_1 = 0.376$, $\zeta_{1.5} = 0.372$, $\zeta_2 = 0.368$, $\zeta_{2.5} = 0.364$ and $\zeta_3 = 0.360$. In this thesis, we use $n = 2$, which gives a polytropic exponent of $\frac{3}{2}$, consistent with protostars in quasi-static equilibrium.

The pseudo-mean optical depth is calculated in the same manner as the pseudo-mean column-density. If we assume that the Rosseland mean opacity is only a function of density and temperature, such that $\kappa(\rho, T)$, then the optical depth from any dimensionless radius to the edge of the pseudocloud is given by

$$\tau_i = \int_{\xi'=\xi}^{\xi'=\xi_B} \kappa_i(\rho_c \theta^n(\xi'), T_c \theta(\xi)) \rho_c \theta^n(\xi') R_0 \, d\xi' \tag{3.63}$$
$$= \left(\frac{-\psi_i \rho_i \theta^n(\xi)}{4\pi \mathrm{G} \phi(\xi)} \right)^{\frac{1}{2}} \int_{\xi'=\xi}^{\xi'=\xi_b} \kappa_i \left(\rho_i \left[\frac{\theta(\xi')}{\theta(\xi)} \right]^n, T_i \left[\frac{\theta(\xi')}{\theta(\xi)} \right] \right) \left[\frac{\theta(\xi')}{\theta(\xi)} \right]^n d\xi'. \tag{3.64}$$

We then take the mass-weighted average, so that we now have, for the mass-weighted pseudo-mean optical depth

$$\bar{\tau_i} = \left(-\xi_B^2 \frac{d\theta}{d\xi}(\xi_b) \right)^{-1} \left(\frac{-\psi_i \rho_i}{4\pi \mathrm{G}} \right)^{\frac{1}{2}} \times \int_{\xi=0}^{\xi=\xi_B} \int_{\xi'=\xi}^{\xi'=\xi_B} \kappa \left(\rho_i \left[\frac{\theta(\xi')}{\theta(\xi)} \right]^n, T_i \left[\frac{\theta(\xi')}{\theta(\xi)} \right] \right) \theta^n(\xi') d\xi' \left[\frac{\theta^n(\xi)}{\phi(\xi)} \right]^{\frac{1}{2}} \xi^2 \, d\xi. \tag{3.65}$$

This is a complicated function, and one that would certainly increase computation time if it had to be evaluated at every timestep. However, we can, fortunately, make use of lookup tables and interpolate regularly to find these values. We define a pseudo-mean mass opacity as

$$\bar{\kappa_i} = \frac{\bar{\tau_i}}{\bar{\Sigma_i}}, \tag{3.66}$$

and the beauty of this is once n is fixed, then $\kappa = f(\rho_i, T_i)$. This needs to be calculated only once, so for a given value of ρ and T, we have

$$\bar{\kappa}(\rho, T) = \left(-\zeta_n \xi_B^2 \frac{d\theta}{d\xi}(\xi_B) \right)^{-1} \int_{\xi=0}^{\xi=\xi_B} \int_{\xi'=\xi}^{\xi'=\xi_B} \kappa \left(\rho \left[\frac{\theta(\xi')}{\theta(\xi)} \right]^n, T \left[\frac{\theta(\xi')}{\theta(\xi)} \right] \right) \theta^n(\xi') d\xi' \left[\frac{\theta^n(\xi)}{\phi(\xi)} \right]^{\frac{1}{2}} \xi^2 d\xi. \tag{3.67}$$

Formally, $\kappa = f(\rho, T)$ only. However, by using the pseudo-mean column density in the calculation of the pseudo-mean optical depth, the local environment heavily influences the optical depth of the particle. In practice, the emission and absorption experienced by the particle is then influenced by the local surroundings, for example, a hot particle may be insulated by cooler surroundings.

This now leads us to the essence of this method, *the polytrope approximation compensates for absorption of escaping radiation from a particle by altering the net radiative loss from the particle.*

Finally, the total cooling term for particle i is given by

$$\left.\frac{du}{dt}\right|_{\rm cool} = \frac{4\sigma(T_0^4(\mathbf{r}) - T_i^4)}{\bar{\Sigma}_i^2 \bar{\kappa}_i(\rho_i, T_i) + \kappa_{{\rm Planck},i}^{-1}(\rho_i, T_i)}, \tag{3.68}$$

where T_0 allows for irradiation from a background source or stellar field, and $\kappa_{{\rm Planck},i}$ is the Planck-mean opacity. We note that Forgan et al. (2009) take the Rosseland-mean and Planck-mean opacities to be equal. The construction of the cooling time allows smooth variation between optically thick and optically thin regimes, and is optimised for optical depths around 1.

Although this method is computationally efficient, its two main limitations are:

1. The pseudo-cloud assumes a spherical polytropic cloud. By definition, this is better suited to spherical objects. For example, if we simulate a fragmenting protostellar disc, this method will best capture the cooling behaviour inside the fragments, where there is a large degree of spherical symmetry. It will not be as accurate in the rest of the disc.

2. Crucially, this method *does not capture the exchange of heat between neighbouring particles.*

Fortunately, flux-limited diffusion (FLD) can capture the necessary physics in item 2 of this list. We now turn our attention to this method.

3.7.2 The Flux-Limited Diffusion Approximation

The FLD method is somewhat less involved than the polytropic cooling approximation. Forgan et al. (2009) based their method on the FLD formalism of Mayer et al. (2007b), which is in turn based on conduction modelling work by Cleary & Monaghan (1999) and the flux-limiter of Bodenheimer et al. (1990). In the diffusion approximation, the energy loss is given by

$$\left.\frac{du_i}{dt}\right|_{\text{diff}} = \sum_b \frac{4m_b}{\rho_i \rho_b} \frac{k_i k_b}{k_i + k_b}(T_i - T_b)\frac{\mathbf{r}_{ib}\cdot\nabla\mathbf{W}}{|r_{ib}^2|}, \tag{3.69}$$

where the sum over b is the neighbours of i, $\mathbf{W}$ is the smoothing kernel, $\mathbf{r}_{ib}$ is the separation vector between i and b, and k_i is the thermal conductivity of the particle, given by

$$k_i = \frac{16\sigma}{\rho_i \kappa_i}\lambda_i T_i^3, \tag{3.70}$$

where k_i is the opacity, σ is the Stefan-Boltzmann constant and λ_i is the flux-limiter, calculated from the local radiation field and given by Bodenheimer et al. (1990) as

$$\lambda_i(R_i) = \frac{2 + R_i}{6 + 3R_i + R_i^2}, \tag{3.71}$$

where R_i is a function of the radiation energy density at the position of particle,$u_r(\mathbf{r}_i)$,

$$R_i = \frac{|\nabla u_r(\mathbf{r}_i)|}{u_r(\mathbf{r}_i)\rho_i\kappa_i}. \tag{3.72}$$

If the region is optically thick, ρ and κ become large, so $R_i \to 0$ and $\lambda_i \to \frac{1}{3}$. In the optically thin regime, R_i becomes large, so $\lambda_i \to 0$, which ends transport by diffusion. As with the polytropic cooling approximation, this method has some

shortcomings, which are

- Radiation is not modelled well at low optical depths, since $\lambda_i \to 0$, which essentially ends transport by diffusion.

- The system cannot lose energy through this prescription, i.e. it cannot radiatively cool. This must be implemented ad hoc assuming prior knowledge of the system.

3.7.3 The Hybrid Radiative Transfer Method

The Forgan et al. (2009) hybrid method blends polytropic cooling with energy transfer from the flux-limited diffusion approximation. In this fashion, the shortcomings of both methods are mitigated, with no overlap. Since FLD cannot model cooling, and the polytropic cooling cannot model energy exchange, combining both methods requires very little additional extra work. The total cooling time for each particle is then simply

$$\left.\frac{du_i}{dt}\right|_{\text{total}} = \left.\frac{du_i}{dt}\right|_{\text{hydro}} + \left.\frac{du_i}{dt}\right|_{\text{cool}} + \left.\frac{du_i}{dt}\right|_{\text{diff}}, \tag{3.73}$$

where $\left.\frac{du_i}{dt}\right|_{\text{hydro}}$ is the energy change to compressive $P\,dV$ hydrodynamic heating.

3.8 Riemann Solvers and Artificial (Thermal) Conductivity

Artificial viscosity is, essentially, a dissipative term added to the SPH equations of energy and motion. The form of the viscosity, given by the properties of μ_{ab} in equation 3.26, was chosen such that it is only present in compression, conserves angular momentum and vanishes in the case of rigid body rotation (and also in shockless differential rotation). However, many astrophysical problems include relativistic outflows, and in such scenarios it was initially unclear what form this viscosity should take.

This motivated the development of viscosity based on Riemann solvers (Chow & Monaghan, 1997; Monaghan, 1997), where the basic idea is to treat interacting particles as the left and right state of the Riemann problem, and, crucially, making the assumption that the jump in characteristics on the left and right of the boundary can be replaced by the difference between the characteristics, considered only along the line of sight between the points. In "standard" SPH viscosity, the dissipative contribution arises from Π_{ab}. Following Monaghan (1997), we now build a form based on the eigenvalue of the Riemann problem, i.e. the propagation speed of the medium, or the *signal velocity*, $v_{\rm sig}$. In "standard" SPH, the dissipative term of the momentum equation is given by

$$\left(\frac{d\mathbf{v}_a}{dt}\right)_{\rm diss} = -\sum_b m_b \Pi_{ab} \nabla_a W_{ab}, \tag{3.74}$$

but in the form based on Riemann solvers we now have

$$\left(\frac{d\mathbf{v}_a}{dt}\right)_{\rm diss} = \sum_b m_b \frac{\alpha v_{\rm sig}\mathbf{v}_{ab}\cdot\hat{\mathbf{e}}_{ab}}{\bar{\rho}_{ab}}\nabla_a W_{ab} \tag{3.75}$$

where $\hat{\mathbf{e}}_{ab}$ is the unit vector of the line joining a and b, $\hat{\mathbf{e}}_{ab} = \frac{\mathbf{v}_{ab}}{|\mathbf{v}_{ab}|}$ and $\mathbf{v}_{ab}\cdot\hat{\mathbf{e}}_{ab}$ is the velocity difference along the line joining a and b. Similarly, for thermokinetic energy, we have

$$\left(\frac{d\hat{e}_a}{dt}\right)_{\rm diss} = \sum_b m_b \frac{e^*_{ab}}{\bar{\rho}_{ab}}\hat{\mathbf{e}}_{ab}\cdot\nabla_a W_{ab} \tag{3.76}$$

where $\hat{e} = \frac{1}{2}v^2 + u$ is the specific thermokinetic energy, and

$$e^*_a = \frac{1}{2}\alpha v_{\rm sig}(\mathbf{v}_a\cdot\hat{\mathbf{e}})^2 + \alpha_u v^u_{\rm sig}u_a, \tag{3.77}$$

which insures positive, definite viscous dissipation (Monaghan, 1997). Here, we have allowed for two signal velocities, and two strengths, $v^u_{\rm sig}$ and α_u is the signal velocity and strength of the thermal energy dissipation, and $v_{\rm sig}$ and α describe the kinetic part. This is important, since it allows us, in the thermal energy equation, to separate the dissipative contribution from the "standard" part of SPH from the smoothing of the discontinuities in thermal energy, and this is known as *artificial (thermal) conductivity*. The full expression for this is

$$\left(\frac{du_a}{dt}\right)_{\rm diss} = -\sum_b \frac{m_b}{\bar{\rho}_{ab}}\left[\frac{\alpha v_{\rm sig}(\mathbf{v}_{ab}\cdot\hat{\mathbf{e}}_{ab})}{2} + \alpha_u v^u_{\rm sig}(u_{ab})\right]\hat{\mathbf{e}}_{ab}\cdot\nabla_a W_{ab}, \tag{3.78}$$

where the second term in this expression is the artificial conductivity, which smooths discontinuities in the thermal energy. We note that this thermal conductivity, which allows us to resolve discontinuities in the thermal energy,

just as artificial viscosity allows us to resolve discontinuities in density, is "*almost universally ignored in SPH formulations*" (Price, 2008).

This forms one of the building blocks into our investigation of the fragmentation boundary in protostellar discs, discussed in section 3.10. Note that $v_{\rm sig}$ does not necessarily need to equal $v^u_{\rm sig}$, which allows individual treatment of kinetic and thermal energy. For non relativistic flows (Price, 2012),

$$v_{\rm sig} = \begin{cases} \frac{1}{2}[c_{s,a} + c_{s,b} - \beta \mathbf{v}\cdot\hat{\mathbf{e}}_{ab}], & \mathbf{v}\cdot\hat{\mathbf{e}}_{ab} \leq 0 \\ 0, & \mathbf{v}\cdot\hat{\mathbf{e}}_{ab} > 0 \end{cases} \tag{3.79}$$

where β acts as the von Neumann and Richtmeyer viscosity. Choosing a form that is 0 if $\mathbf{v}\cdot\hat{\mathbf{e}}_{ab} > 0$ ensures that the dissipative action only occurs for approaching particles. With this form, the signal velocity is largest when the velocity vector of each particle only has non-zero components along the line of sight joining the two particles.

The form of the signal velocity for the thermal contribution is of particular interest to us, since we have employed this additional dissipative term in our investigation of the fragmentation boundary. In shock regions, the signal velocity described in equation 3.79 is appropriate, since shocks travel at sound speed and undergo compression. However, Price (2008) realised that this may not be suitable for discontinuities, by considering two regions of different temperature and density, but in pressure equilibrium. If thermal conductivity with a signal velocity based on sound speed is used, the region would undergo diffusion until no temperature gradient remains. Instead, the author proposed a thermal signal velocity of the

form

$$v_{\rm sig}^{u} = \sqrt{\frac{|P_a - P_b|}{\bar{\rho}_{ab}}}, \tag{3.80}$$

which we can see vanishes if the pressure in the two regions is equal. It effectively treats discontinuities in the thermal energy at pressure boundaries without introducing unwanted dissipation elsewhere. Unfortunately, this particular form of the signal velocity is unsuitable for situations where self-gravity is present. This is because in the case of hydrostatic equilibrium, the gravitational force is balanced by a pressure gradient force, which means that the artificial conductivity term will not switch off, even though the medium is in equilibrium. Instead, in this work, we use the form suggested by Wadsley et al. (2008), which in terms of this artificial conductivity is equivalent to (Price, 2008)

$$v_{\rm sig}^{u} = |\mathbf{v}_{ab} \cdot \hat{\mathbf{e}}_{ab}|. \tag{3.81}$$

We can see that this form will always be non-zero so long as there is shear flow, i.e. off in the case of solid body rotation, and on in the case of shear flow (i.e. differential rotation). Since our aim here is to investigate the non-convergent behaviour of the fragmentation boundary of differentially rotating protostellar discs, this uniformly acting conductivity is expected to influence the fragmentation boundary of our discs.

This has brought us, nicely, up to speed with SPH, enough to formulate our problem, and our initial investigations toward a solution.

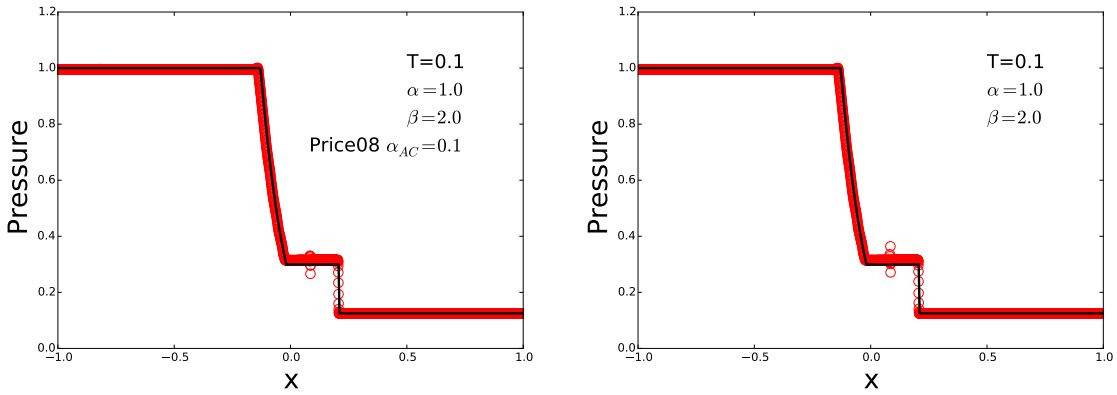


Figure 3.6: Pressure term in shock tube test, left panel shows the implementation of artificial conductivity, acting to smooth out the pressure "blip". Right hand side shows the same SPH implementation but without the thermal conductivity term. Note the use of a small amount of thermal conductivity here. Since we used $\alpha = 0.1$ and $\beta = 0.2$ for viscosity in our convergence testing described later in this chapter, we set the strength of artificial conductivity to be comparable, i.e. $\alpha_{\rm AC} = 0.1$. Using the same low level of viscosity in a low resolution shock tube simulation, however, results in interparticle penetration.

3.9 Author's Note

The following section contains unpublished work that was carried out in the first $\sim$ 6 months of my PhD. It did not lead to a decisive conclusion to our investigation, but this, in hindsight, is unsurprising, and is discussed in our conclusions for this chapter in section 3.11.

3.10 Convergence

3.10.1 Introduction

As explained in the introduction to this thesis, accretion discs are ubiquitous in astronomy and can be found at many scales, from black holes to protostars. If

the disc around the central object is sufficiently massive, its own self-gravity may play an important role in its evolution, through the growth of the gravitational instability (Goldreich & Lynden-Bell, 1965; Safronov, 1960). The susceptibility of self-gravitating protostellar discs to gravitational instability is characterised by the Toomre parameter Q (Toomre, 1964)

$$Q = \frac{c_s \kappa}{\pi G \Sigma}, \tag{3.82}$$

where c_s is the sound speed in the disc, Σ is the surface density of the disc and κ is the epicyclic frequency, which is the frequency at which a radially displaced particle of fluid will oscillate. For Keplerian discs, this can be replaced with the angular frequency Ω. Discs are unstable to axisymmetric perturbations for $Q < 1$ (Toomre, 1964), and modern numerical simulations show that discs are unstable to nonaxisymmetric perturbations for $Q \lesssim 1.5$ (Durisen et al., 2007). As these spiral waves dissipate, they provide a source of heat which can balance cooling such that these discs may exist in a state of marginal stability (Paczynski, 1978). In these marginally stable discs, the instability acts to transport angular momentum outwards, allowing mass to accrete onto the central object (Laughlin & Bodenheimer, 1994; Lin & Pringle, 1987; Lodato & Rice, 2004). If the cooling time in the disc is short, then the disc may fragment to form gravitationally bound objects, such as giant planets in a protostellar disc (Boss, 1998, 2000), or stars around a supermassive black hole (Bonnell & Rice, 2008).

It was later shown that in addition to the Toomre (1964) parameter the evolution of a self-gravitating disc is also controlled by the rate at which it loses energy,

characterised by cooling time τ_c (Gammie, 2001)

$$\tau_c = \frac{t_{\rm cool}}{t_{\rm dynamical}} \frac{1}{\Omega} = \frac{\beta}{\Omega}, \tag{3.83}$$

where β is some dimensionless constant. Gammie (2001) showed in 2D shearing sheet simulations that for $\beta < \beta_{\rm crit}$, with $\beta = 3$ and specific heat ratio $\gamma = 2$, that the disc would fragment, and for $\beta > \beta_{\rm crit}$ the disc would settle into a quasi-equilibrium state. Simply put, for very short cooling times, regions in the disc collapse before they are able to collide with other regions and reheat. Results consistent with this were found by Rice et al. (2003b), although the actual value of $\beta_{\rm crit}$ was found to depend upon the equation of state, γ (Rice et al., 2005b).

As previously mentioned, accretion discs act to transport angular momentum outwards. Doing so requires a torque, Γ, and the most simple assumption is that this originates from a fluid viscosity ν. However, the molecular fluid viscosity, for a typical protostellar disc, gives a viscous timescale that is longer than the age of the universe[17]. Therefore, if the torque redistributing angular momentum is viscous in origin, it must come from some additional, turbulent viscosity. Assuming that the disc *is* turbulent (of which there is, to the author's knowledge, no proof, to date), then considering that $\nu = vl$, where ν is the viscosity, v is the characteristic velocity and l is the characteristic length that characterises the turbulence, Shakura & Sunyaev (1973) argued that the viscosity could be parameterised using the characteristic velocity of the disc, the sound speed c_s,

[17] $r \sim 10^{14}$ cm, $v_{therm} \sim 10^5$ cm s^{-1}, mean free path $\lambda \sim 10$ cm, $\nu = \lambda v_{therm} = 10^6$ cm^2 s^{-1}, $\tau = \frac{r^2}{\nu} \sim 10^{14}$ years.

and the characteristic length, H, such that

$$\nu = \alpha c_s H, \tag{3.84}$$

where α is a dimensionless parameter, both characterising the strength of the viscosity and absorbing our uncertainty. Gammie (2001) showed that the viscosity could be related to the (local) cooling time, τ_c, of the disc by

$$\alpha = \frac{4}{9\gamma(\gamma-1)\tau_c\Omega}, \tag{3.85}$$

where γ is the ratio of specific heats and Ω is the angular velocity. Rice et al. (2005b) showed that the fragmentation boundary did not only depend upon τ_c, but also on γ. This means that, actually, the fragmentation boundary depends upon the stress in the disc, which is characterised by α. Physically, α is a measurement of the rate of energy extraction from the shear flow, which is then dumped into the system as heat. High values of α correspond to rapid extraction of energy. If energy is extracted too rapidly, then regions in the disc will collapse, since they will not have a chance to collide and reheat the disc. This corresponds to a maximum stress that the disc can sustain without fragmenting, Gammie (2001) measured $\alpha \sim 0.02$ in a 2D shearing box, and Rice et al. (2005a) found results consistent with this, $\alpha \sim 0.06$, in 3D global simulations of fragmenting protostellar discs. From here, the basic picture began to develop that the perturbation amplitudes of the spiral arms will depend on the cooling rate, since it was found that the perturbation amplitude, $^{\delta\Sigma}/_{\Sigma}$, is related to the

cooling parameter β by (Cossins et al., 2009)

$$\frac{\delta\Sigma}{\Sigma} \sim \frac{1}{\sqrt{\beta}}, \tag{3.86}$$

and similarly (Rice et al., 2011)

$$\frac{\delta\Sigma}{\Sigma} \sim \alpha^{\frac{1}{2}}. \tag{3.87}$$

Essentially, this means that there is a maximum perturbation amplitude that the disc can sustain in its spiral arms, without the effective stress, α (which we can also think of the rate of extraction of energy), becoming so large that the disc fragments.

However, recent SPH simulations by Meru & Bate (2011) that fix β show that as resolution increases, the value of β for which fragmentation will occur also increases. If this is physical, rather than numerical, then the implication is that fragmentation can occur for arbitrarily long cooling times. Although this means that fragmentation could happen in the inner regions of protostellar discs, forming gas giants, it also means that fragmentation will take place for arbitrarily small amplitudes and stresses. If convergence of this value is never achieved, the implication is that all discs, given enough time, will eventually be susceptible to fragmentation. If, however, Meru & Bate (2011) are correct, it means that we may need to rethink our entire understanding of self-gravitating accretion discs.

More recently, Meru & Bate (2012) showed that convergence *does* occur when the strength of the artificial viscosity in the SPH simulation is adjusted. They

suggest that with typical artificial viscosity parameters from past SPH studies ($\alpha_{\rm SPH} = 0.1$, $\beta_{\rm SPH} = 0.2$), the simulations are converging on a value of $\beta = 17.4$, but with different parameters ($\alpha_{\rm SPH} = 0.1$, $\beta_{\rm SPH} = 2.0$), the converging value is $\beta = 30$. Typically, SPH simulations use $\beta_{\rm SPH} = 2\alpha_{\rm SPH}$ so that the linear term dominates at slow particle approach, and in the region of shocks (i.e. fast particle approach) the quadratic term dominates. By using $\beta_{\rm SPH} = 20\alpha_{\rm SPH}$, the quadratic term, effectively, always dominates. It is difficult to see how this will avoid changing the simulation properties.

At the start of CH's PhD (September 2013), a robust solution to this problem had not yet been found. It had been suggested (Lodato & Clarke, 2011) that numerical viscosity may, in fact, be more important than had previously been thought, requiring higher resolutions than has previously been thought to correctly resolve fragmentation. Rice et al. (2012b) suggested that this lack of convergence is directly linked to the implementation of β cooling. In SPH, physical quantities are determined by density weighted interpolations. However, by implementing a cooling law of the form

$$\left(\frac{du_j}{dt}\right)_{\rm cool} = -\frac{u_j}{\tau_c}, \tag{3.88}$$

this change in thermal energy occurs only at the location of the particle, i.e. it is not smoothed amongst its neighbours, nor convolved with a smoothing kernel. Instead, Rice et al. (2012b) proposed a new type of cooling, so-called "smoothed cooling", of the form

$$\left(\frac{du_j}{dt}\right)_{\rm cool} = -\frac{1}{\tau_{\rm c}}\sum_i \frac{m_j}{\rho_j} u_i W(\mathbf{r}_{ji}, h), \tag{3.89}$$

where the sum over i is over the particles in the neighbour sphere of j. If $i = j$, then the kernel becomes $W(0, h_j)$. Performing SPH simulations with this smoothed cooling implemented initially appeared to solve the convergence problem, up to a resolution of 2 million particles, as Rice et al. (2012b) noted. However, their 10 million particle run failed to converge even with this smooth cooling. We note now, for later discussion, that this 10 million particle run was an "effective" 10 million particle run. That is, a run using 4 million particles located in a ring, to capture global gravitational effects, while reducing computational requirements.

This is where CH began this work, and it is now outlined. As noted, the work that follows did not lead to a decisive conclusion to our investigation, but we discuss the reasons for this, in relation to the work conducted by Rice et al. (2014a), in the conclusion of this chapter.

$$\frac{v_\phi^2}{r} = \frac{\partial \Phi}{\partial r}, \tag{3.90}$$

3.10.2 Artificial Conductivity. A Solution to Fragmentation due to Numerical Effects?

The first investigation into convergence was performed using artificial conductivity of the form described in Wadsley et al. (2008), and outlined earlier in section 3.8.

Our rationale was based upon the ability of thermal conductivity to smooth pressure discontinuities in Sod shock tube problems. Indeed, since Rice et al.

(2012b) noted that smoothed cooling removes this discontinuity, the similarity between the effects of the two was what drove this investigation. As mentioned earlier, we use a signal velocity of the form Wadsley et al. (2008), since in self-gravitating simulations, where in hydrostatic equilibirum the gravitational force is supported by the pressure, the artificial conductivity given in Price (2008) would cause unphysical cooling.

3.10.2.1 Suite of Simulations to Investigate the Effect of Artificial Conductivity on Convergence of the Fragmentation Boundary

We present a suite of 102 global SPH simulations conducted using the `SEREN` SPH code (Hubber et al., 2011), at four different resolutions (250,000 particles, 500,000 particles, 1 million particles and 2 million particles). At 2.5×10^5, 5.0×10^5 and 1.0×10^6 particle resolution, 3 different values of β (where $\beta = t_c \Omega$) are selected, and for each value of β, 5 discs are run, which are identical apart from the random number seed used to initialise the disc. For each of these, we run the simulations with and without artificial conductivity turned on. This totals 30 simulations for each resolution at 2.5×10^5, 5.0×10^5 and 1.0×10^6. The effect of artificial conductivity in halting fragmentation is shown in Figure 3.7, which displays two SPH discs of 1 million particles and $\beta = 8$. The disc on the left hand side was run without artificial conductivity, and fragmented. The disc on the right hand side is the exact same disc, run with a small amount of artificial conductivity. It has not fragmented.

Due to computational limitations, for the 2 million particle runs, we run only 12 simulations at this resolution. The details of each simulation are outlined in Table

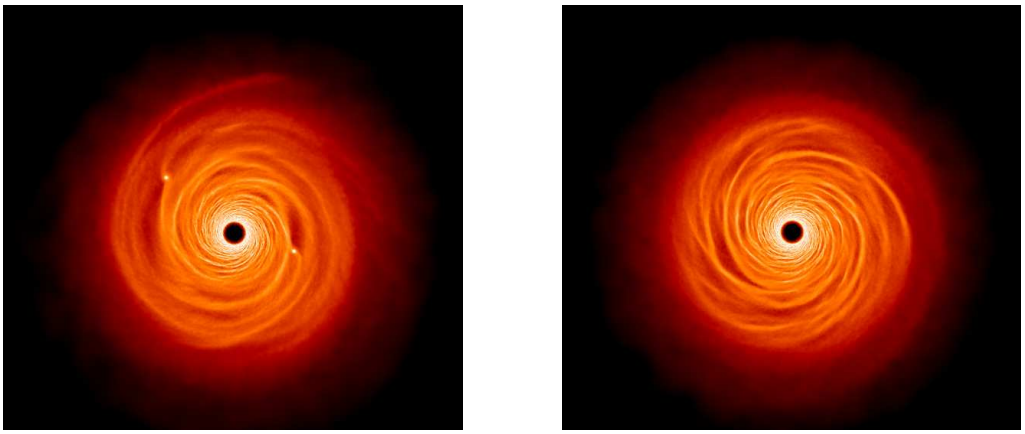


Figure 3.7: Both discs contain 1 million particles and have $\beta = 8$. The disc on the left has no artificial conductivity, and the disc on the right has a small amount of artificial conductivity ($\alpha_u \approx 0.1$ in equation (3.78)). The left disc has undergone fragmentation whereas the right one, with artificial conductivity, has not.

3.1. We select β values that increase with resolution, since we are investigating the convergence of the fragmentation boundary. The constant determining the strength of the artificial conductivity was chosen as $\alpha_{\rm cond} = 0.1$ since the $\alpha_{\rm SPH}\beta_{\rm SPH}$ formalism of SPH was used where $\alpha_{\rm SPH} = 0.1$ and $\beta_{\rm SPH} = 0.2$, so the thinking was that the strength scale would be similar. As this is the first time this is investigated, it was unclear if this was a sensible selection at the time. Simulations that fragment are marked with a tick, those that do not are marked with a cross. Simulations that are not run due to computational expense are marked with a hyphen.

The results of this experiment are shown in Table 3.1. From left to right, the columns are resolution, β, and at each disc if it was run without artificial conductivity (NAC) or with artificial conductivity (AC0.1).

$\mathbf{N \times 10^6}$	β	**Disc 1**		**Disc 2**		**Disc 3**		**Disc 4**		**Disc 5**	
		NAC	AC0.1	NAC	AC0.1	NAC	AC0.1	NAC	AC0.1	NAC	AC0.1
	5	✓	✓	×	✓	✓	×	✓	✓	✓	×
0.25	6	×	✓	×	×	×	×	×	×	×	×
	7	×	×	×	×	×	×	×	×	×	×
	5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
0.5	6	×	×	✓	✓	✓	✓	✓	✓	✓	×
	7	×	×	×	×	×	×	×	×	×	×
	7	✓	✓	✓	✓	✓	×	✓	✓	✓	✓
1.0	8	✓	×	✓	×	✓	×	×	×	×	×
	9	×	×	×	×	×	×	✓	×	×	×
	7	✓	✓	✓	-	-	-	-	-	-	-
	8	✓	✓	✓	-	-	-	-	-	-	-
2.0	9	×	×	✓	-	✓	✓	✓	✓	✓	✓
	10	-	-	-	-	-	-	-	-	-	-
	11	-	-	-	-	-	-	-	-	-	-
	12	-	-	-	-	-	-	-	-	-	-

Table 3.1: This table shows simulations of varying resolution (number of particles) and varying β. Ticks indicate that the discs underwent fragmentation, whilst crosses indicate that they did not. Hyphens indicate simulations in progress. It was originally thought that the addition of an artificial conductivity term may result in the code converging for fragmentation with increasing resolution, but this does not robustly appear to be the case. For lower amounts of AC, it will reduce the amount of fragmentation. For too large an amount, more fragments appear which is most likely due to interference with shocks.

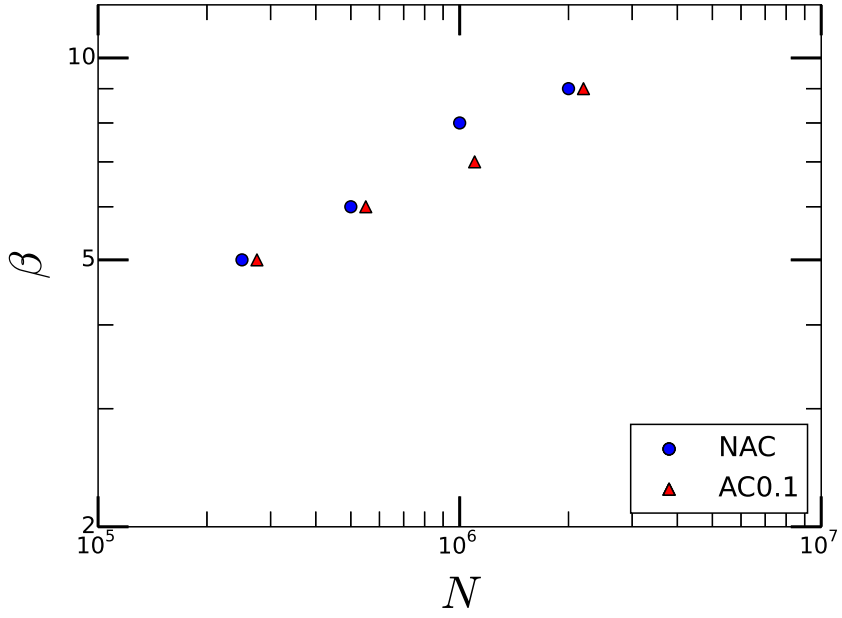


Figure 3.8: Plot showing β at which fragmentation occurred as a function of resolution, where N is the number of particles in the SPH simulation. Blue circles show simulations without artificial conductivity, red triangles show simulations that included artificial conductivity with $\alpha_{\rm cond} = 0.1$. Red triangles, depicting runs where $\alpha_{\rm cond} = 0.1$, are offset by +10% on the x-axis for clarity.

For clarity, the results shown in Table 3.1 are plotted in Figure 3.8, which shows the last value of β for which fragmentation will occur, as a function of resolution, N, where N is the number of SPH particles in the simulation. Blue circles indicate runs with no artificial conductivity, red triangles mark runs where artificial conductivity is included, with $\alpha_{\rm cond} = 0.1$. The last value of β for which fragmentation occurs is determined by at least 3 out of the 5 simulations tabulated in Table 3.1 fragmenting. Red triangles, depicting runs with artificial conductivity, are offset by +10% on the x-axis for clarity.

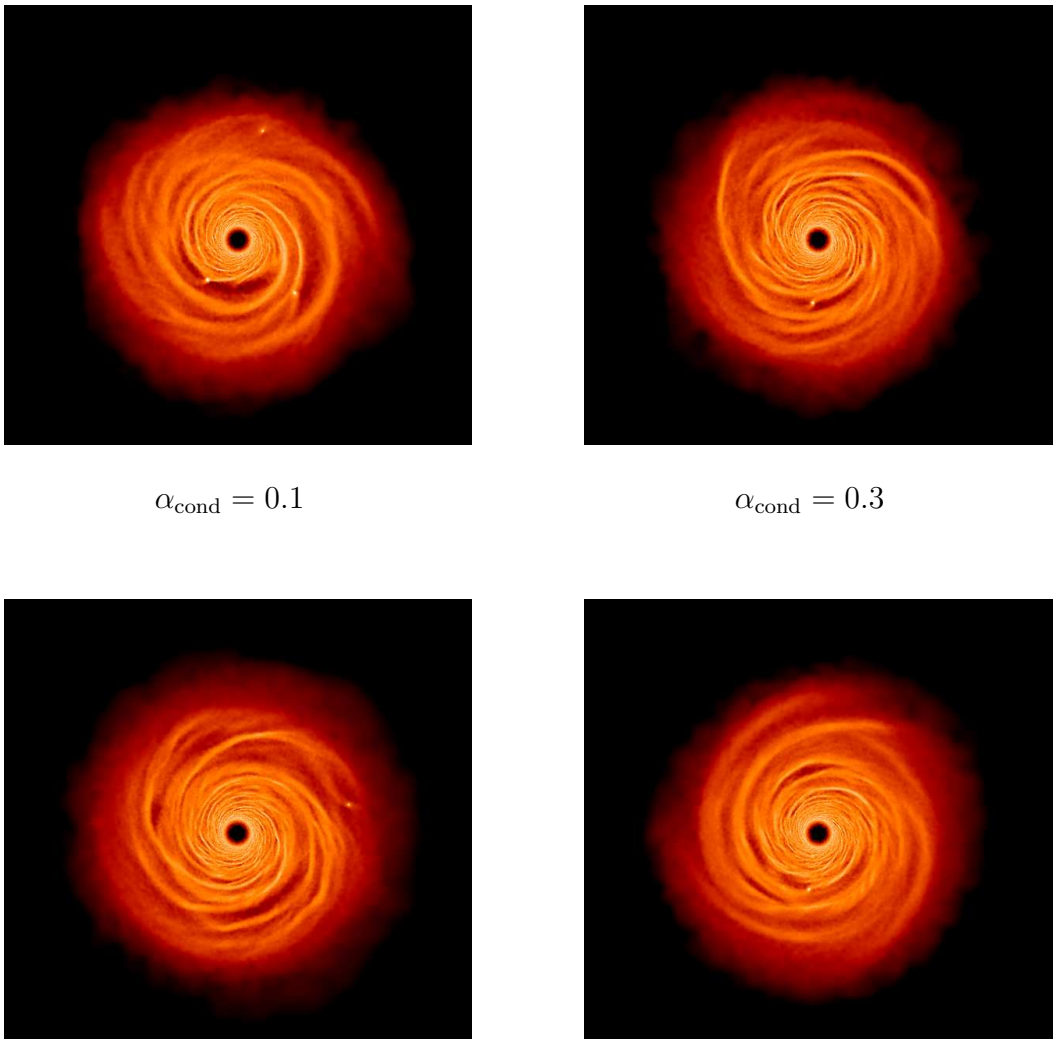
Table 3.1 and Figure 3.8 show an interesting result. For lower resolution this artificial conductivity has very little effect, but at a resolution of 1 million it

prevents fragmentation for several discs. At a resolution of 2 million, however, this does not appear to be the case. It is unclear whether the constant determining the required strength of the artificial conductivity has a more complicated relationship than "about the same as the artificial viscosity". For this reason, we ran an additional suite of simulations at four resolutions, using discs that fragmented with $\alpha_{\rm cond} = 0.1$. This is shown in Table 3.2, which shows 17 simulations, at different resolutions, with different artificial conductivity strengths. Ticks mean the disc fragmented, crosses mean it did not, and circles mean the simulation was not run due to limited computational resources.

$\mathbf{N \times 10^6}$	β	$\mathbf{AC = 0.1}$	$\mathbf{AC = 0.3}$	$\mathbf{AC = 0.5}$	$\mathbf{AC = 1.0}$	$\mathbf{AC = 2.0}$
0.25	5	✓	✓	✓	✓	o
0.5	6	✓	✓	✓	✓	o
1.0	7	✓	✓	✓	×	o
2.0	8	✓	✓	✓	✓	✓

Table 3.2: This table shows discs which fragment when $\alpha_{\rm cond} = 0.1$. We re-ran this discs with increased $\alpha_{\rm cond}$, at 0.3, 0.5, 1.0 and 2.0. Ticks mean the disc fragmented, crosses mean it did not. A circle means the simulation was not run due to limited resources. Although it does not really appear to alter the absolute boundary at which fragmentation occurs, these discs do look different.

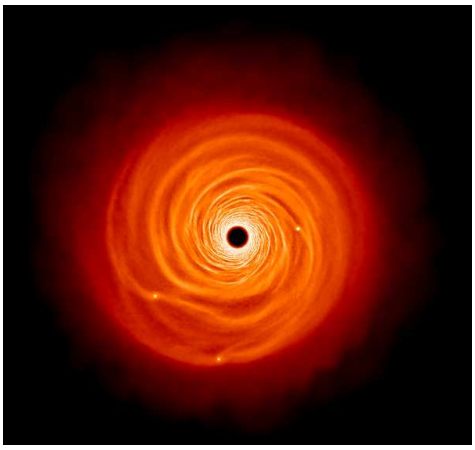
We can see that changing the strength of the artificial conductivity did not prevent fragmentation of the disc in any case except for the $N = 1.0 \times 10^6$ particle



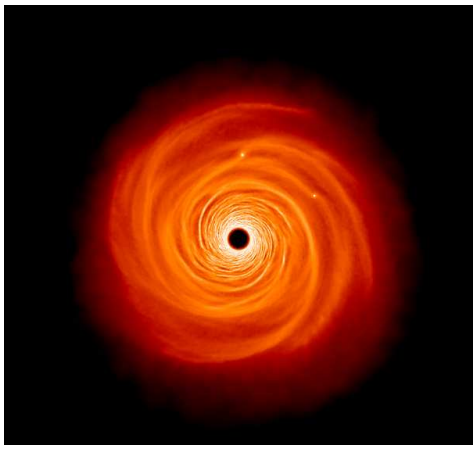
$\alpha_{\rm cond} = 0.1$ $\alpha_{\rm cond} = 0.3$

$\alpha_{\rm cond} = 0.5$ $\alpha_{\rm cond} = 1.0$

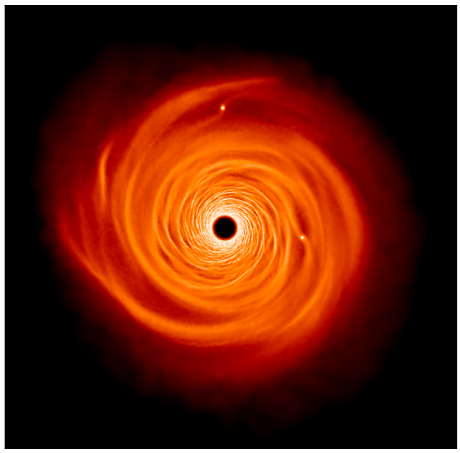
Figure 3.9: This is a table of discs at a resolution of 0.25 million particles, with $\beta = 5$ and varying strengths of artificial conductivity. This value of β was selected because it was the borderline value for fragmentation. Different amounts of artificial conductivity changes how the disc appears.



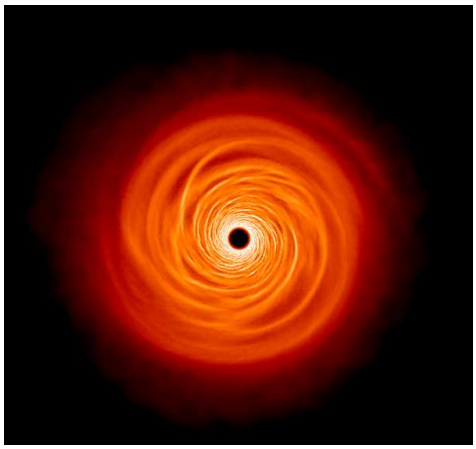
$\alpha_{\rm cond} = 0.1$



$\alpha_{\rm cond} = 0.3$



$\alpha_{\rm cond} = 0.5$



$\alpha_{\rm cond} = 1.0$

Figure 3.10: This is a table of discs at a resolution of 1 million particles, with $\beta = 7$ and varying strengths of artificial conductivity. This value of β was selected because it was the borderline value for fragmentation. The difference in the amount of artificial conductivity changes the appearance of the discs, and for a sufficient strength halts fragmentation altogether.

case, where using $\alpha_{\rm cond} = 1.0$ prevented fragmentation. In general, increasing the amount of artificial conductivity reduces the number of fragments that form in any given simulation, but the prevention of fragmentation appears to have an element of stochasticity. We show, in Figures 3.9 and 3.10, how artificial conductivity may reduce or prevent fragmentation.

3.10.3 The Importance of Choice of SPH Implementation in Fragmentation Convergence Testing

During our prior investigations, the question arose as to whether the traditional, standard implementation of SPH, or perhaps the implementation of viscosity, was at the root of the problem. To investigate this, we ran global convergence tests of the fragmentation boundary using three different implementations of SPH, the "standard" Monaghan (1992), the conservative "grad-h" approach (Price & Monaghan, 2004; Springel & Hernquist, 2002) and the pressure based implementation of Ritchie & Thomas (2001). We run simulations using each of these implementations, both with the standard $\alpha\beta$ implementation of artificial viscosity and with Monaghan (1997) artificial viscosity, which is based on Riemann solvers.

3.10.3.1 Standard SPH

We have already discussed, at length, the "standard" implementation of SPH. We just note here that in the standard implementation of SPH (see,e.g., Monaghan 1992), we assume that the density across the smoothing kernel is constant, and

that the smoothing length for each particle is usually

$$h_i \propto \left(\frac{1}{\rho_i}\right)^{\frac{1}{D}}, \tag{3.91}$$

but since ρ_i is a function of h_i, then we solve

$$\frac{dh_i}{dt} = -\frac{h_i}{D\rho_i}\frac{d\rho_i}{dt}, \tag{3.92}$$

alongside the other physical properties, and for interacting particles we take the arithmetic mean ($h_{ab} = \frac{1}{2}(h_a + h_b)$).

3.10.3.2 Ritchie-Thomas SPH

Standard implementations of SPH have numerical problems in regions with large density contrasts. Of particular interest is the difficulty created when a cold clump of gas is located in a warmer medium. If large density contrasts are present between the cold clump and the warmer gas, then standard SPH will over-estimate the density of the warmer gas, resulting in the gas rapidly cooling and accreting onto the clump (Pearce et al., 1999). Our line of reasoning here was that, perhaps, the lack of convergence could be due to artificially accelerated cooling of the fragments at lower resolution, and so we decided to employ this implementation of SPH.

Instead of assuming constant density over the smoothing kernel, Ritchie & Thomas (2001) assume constant pressure instead. To get a physical property, we sum over pressure for each particle, and then use the equation of state to

calculate the local density. For an equation of state given by

$$P = \frac{2}{3}\rho\epsilon, \tag{3.93}$$

where P is pressure, ρ is density and the quantity ϵ is given by

$$\epsilon = \frac{3}{2}\frac{\mathrm{k_b}T}{\mu m_\mathrm{H}}, \tag{3.94}$$

where $\mathrm{k_B}$ is the Boltzmann constant, T is temperature, m_H is the mass of hydrogen and μ is the relative molecular mass (0.6). Using equation 3.93, and using $\rho_i = \sum_j m_j W_{ij}$ we then have the following, interpolated value for pressure

$$P_i = \frac{2}{3}\sum_j m_j \epsilon_j W_{ij}, \tag{3.95}$$

from which the density is then simply

$$\rho_i = \frac{3P_i}{2\epsilon_i}. \tag{3.96}$$

For the smoothing length, h, rather than keeping the number of neighbours approximately constant, with the smoothing length guessed from the previous timestep such that

$$h_{i+1} = h_i\left[\alpha + (1-\alpha)\left(\frac{N_\mathrm{SPH}}{N_i}\right)^{\frac{1}{D}}\right], \tag{3.97}$$

where α is a convergence parameter (typically ~ 0.4, but must be < 1 to avoid overshooting in regions of density contrast), N_SPH is the desired number

of neighbours (typically 32) and N_i is the actual number of neighbours in the neighbour sphere, Ritchie & Thomas (2001) instead weight the particles in the neighbour sphere, such that

$$N_i = \sum_j \frac{2\bar{\rho}_i}{\bar{\rho}_i + \rho_j}. \tag{3.98}$$

This prevents N_i from oscillating between too low and too high in regions of high density contrast.

3.10.3.3 Grad-h SPH

Finally, we are brought to the fully conservative grad-h formalism of SPH, discussed in Springel & Hernquist (2002) and Price & Monaghan (2004), and so-called because of the additional ∇h terms necessary in the SPH equations of motion in order to be fully self-consistent. It has been known for a long time that errors of $\sim 10\%$ in the conservation of energy (or entropy) are possible when using variable smoothing smoothing lengths (for extreme cases of adiabatic shock) (Hernquist, 1993). The inclusion of grad-h terms in the SPH equations of motion was first achieved by Nelson & Papaloizou (1994). However, this was difficult to implement, requiring that the smoothing length be an explicit function of the distance between particles, and that the number of neighbours remains roughly constant. Later, Springel & Hernquist (2002) showed that the grad-h terms could be included directly in the equations of motion by using the Lagrangian of the system and making use of Lagrange multipliers. If the smoothing length is

expressed as a function of density, then for $h = h(\rho)$ (Monaghan, 2002)

$$\rho_i = \sum_j m_j W(\mathbf{r}_{ij}, h_{ij}), \tag{3.99}$$

we take the time derivative

$$\frac{d\rho_i}{dt} = \frac{1}{\Omega_i} \sum_j m_j \mathbf{v}_{ij} \cdot \nabla_i W_{ij}(h_i) \tag{3.100}$$

where

$$\Omega_i = \left[1 + \frac{\partial h_i}{\partial \rho_i} \sum_j m_j \frac{\partial W_{ij}(h_i)}{\partial h_i} \right]^{-1}. \tag{3.101}$$

The momentum equation is then

$$\frac{d\mathbf{v}_i}{dt} = -\sum_j m_j \left[\frac{P_i}{\Omega_i \rho_i^2} \nabla_i W_{ij}(h_i) + \frac{P_j}{\Omega_j \rho_j^2} \nabla_i W_{ij}(h_j) \right] \tag{3.102}$$

and the energy equation is

$$\frac{du_i}{dt} = \frac{P_i}{\Omega_i \rho_i^2} \sum_j m_j \mathbf{v}_{ij} \cdot \nabla_i W_{ij}(h_i), \tag{3.103}$$

as given by Price & Monaghan (2004), but we also refer the reader to Price (2008) for a comprehensive review and a full derivation of these equations beginning from the Lagrangian. Artificial viscosity is added, as in equations 3.24 and 3.27, through the use of Π_{ij}. We note that if the smoothing length is of the form

$$h_i \propto \left(\frac{1}{\rho_i} \right)^{\frac{1}{D}}, \tag{3.104}$$

then since

$$\rho_i = \sum_j m_j W(\mathbf{r}_{ij}, h_i), \tag{3.105}$$

we can see that equations 3.104 and 3.105 must be solved simultaneously, usually iteratively through the Newton-Raphson method or by using Bisection.

Once again, our line of reasoning was that since we were faced with a convergence problem, and knowing that the standard implementation of SPH is not fully conservative without these additional grad-h terms, that performing the convergence tests with the particular implementation of SPH may resolve the issue.

3.10.3.4 Monaghan (1997) viscosity

In section 3.8, we discussed the treatment of viscosity by analogy to the Riemann problem, so we will not repeat what we have already said. We just remind the reader that the pairwise viscosity between two particles can be treated as left and right states of a discontinuity. Dissipative terms are moderated by a signal velocity, and the jump across variables is assumed to be the velocity difference between the two particles along the line of sight joining them ($\mathbf{v}_{ij} \cdot \hat{\mathbf{e}}_{ij}$).

3.10.3.5 Results

We now present a suite of 71 SPH simulations, with three different implementations of SPH, (standard, grad-h and Ritchie & Thomas 2001), each with two viscosity implementations, the standard $\alpha\beta$ approach, and the Monaghan (1997) approach inspired by Riemann solvers. We detail simulations conducted at 5 resolutions, where the number of particles, N, are $N = 2.5 \times 10^5$, $N = 5.0 \times 10^5$,

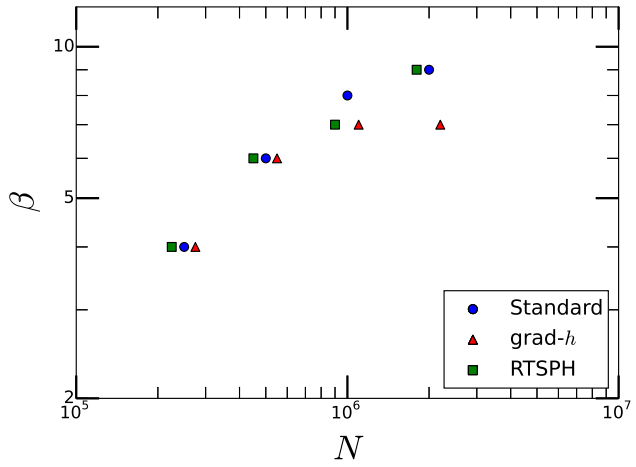


Figure 3.11: Plot showing β at which fragmentation occurred as a function of resolution, where N is the number of particles in the SPH simulation. Blue circles show simulations evolved with standard SPH, red triangles show simulations evolved with grad-h SPH, and green squares show simulations evolved with Ritchie & Thomas (2001) SPH. Grad-h SPH points, marked by red triangles, are offset by +10% on the x-axis. Ritchie & Thomas (2001) points, marked by green squares, are offset by −10% on the x-axis. Both offsets are done for clarity. We only include points up to 2 million particles since not all types of SPH were run for a resolution of $N = 4\times10^6$. Although at 2 million particles, it appears that convergence has been reached, we need to simulate these discs at higher resolution, $N \sim 10^7$, perhaps, to confirm this. However, it is just one result, and may well simply reflect an element of the stochastic nature of fragmentation.

$N = 1.0 \times 10^6$, $N = 2.0 \times 10^6$ and $N = 4.0 \times 10^6$. The results are given in Table 3.3, ticks mean the simulation fragmented, crosses mean it did not fragment, open circles mean the simulation was not run due to consideration for computational resources. We stress that at each resolution, the *same* initial disc was used. Any resulting differences are therefore due to the simulation, and not some slightly different initial condition.

The results in Table 3.3 are also shown in Figure 3.11, depicting the last β at which fragmentation occurred as a function of resolution, N. Blue circles show

simulations evolved with standard SPH, red triangles show simulations evolved with grad-h SPH, and green squares show simulations evolved with Ritchie & Thomas (2001) SPH. Grad-h SPH points, marked by red triangles, are offset by $+10\%$ on the x-axis. Ritchie & Thomas (2001) points, marked by green squares, are offset by -10% on the x-axis. Both offsets are done for clarity. We only include points up to 2 million particles since not all types of SPH were run for a resolution of $N = 4 \times 10^6$. We only include standard viscosity, since grad-h required very low β to fragment at all.

Upon examining Table 3.3 and Figure 3.11, we can see that there was no "goldilocks" solution for which convergence of the fragmentation boundary occurred. Tentatively, at a resolution of 2 million particles and $\beta = 8$, we see that the standard SPH simulation fragmented while the grad-h simulation did not. However, this is just one result, and probably in part reflects the somewhat stochastic nature of fragmentation.

$\mathbf{N \times 10^6}$	β	**Standard**		**Grad-h**		**RTSPH**	
		$\alpha\beta$	M97	$\alpha\beta$	M97	$\alpha\beta$	M97
	0.1	o	o	o	✓	o	o
	3	o	o	✓	×	✓	✓
	4	✓	✓	✓	×	✓	✓
0.25	5	×	×	×	×	o	o
	6	×	×	×	×	o	o
	7	×	×	×	×	o	o
	3	o	o	o	×	o	o
	4	o	o	o	×	o	o
0.5	5	✓	✓	✓	×	✓	✓
	6	✓	✓	✓	×	✓1	×
	7	×	×	×	×	o	o
	7	✓	✓	✓	×	✓	✓
1.0	8	✓	✓1	×	o	×	×
	9	×	o	o	o	o	o
	7	✓	o	-	o	o	o
2.0	8	✓	✓	×	×	✓	✓
	9	✓	o	o	o	✓	✓
	8	✓	o	-	o	o	o
4.0	9	✓	o	-	o	o	o
	10	o	o	o	o	-	o

Table 3.3: Table of convergence, demonstrating three different implementations of smoothed particle hydrodynamics. Standard, conservative "grad-h" and Ritchie Thomas. These are combined with two different implementations of artificial viscosity, the standard $\alpha\beta$ and the Monaghan (1997) viscosity, based on Riemann solvers. Ticks mean the disc fragmented, crosses mean it did not. Circles mean that disc was not run due to computational limitations. Hyphens indicate that the disc did not visibly fragment, however the simulation was unable to run further due to incredibly small timesteps. This indicates a region of very high density, as such, these discs probably will undergo fragmentation.

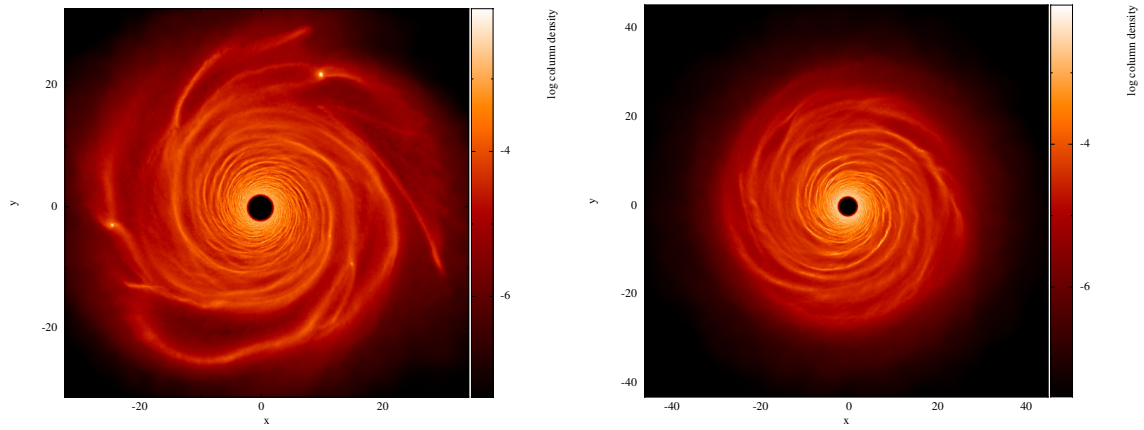


Figure 3.12: Both panels show a 2 million particle disc, simulated with $\beta = 8$. Left is with standard SPH, and standard $\alpha\beta$ viscosity, and it has fragmented at $T = 3000$ years. Right is the same disc, but simulated using grad-h SPH, and the same standard $\alpha\beta$ viscosity. It has not fragmented to form bound clumps, at $T = 7000$ years.

For the conservative grad-h formalism of SPH coupled with Monaghan (1997) viscosity, initially none of the discs fragmented, despite having cooling times rapid enough to do so. It took reducing the β to 0.1 before the disc fragmented. This suggests that using these two options together is causing shocks to be completely smeared out, although exactly *why* this is happening still evades us. One possibility is that by treating interacting particles as left and right states of the Riemann problem, and using the fully conservative grad-h, we have increased the level of mixing between cool, collapsing regions and warmer regions in the simulation to a point where fragmentation is artificially suppressed. Whether or not this explanation is correct requires further investigation, but, regardless of its origin, the result is certainly interesting, and may suggest that there is a problem when combining grad-h SPH with artificial viscosity based on Riemann solvers.

Our next point of interest is that using different SPH implementations can change whether or not the disc fragments. For example, Figure 3.12 shows two discs,

simulated with 2 million particles, both with $\beta = 8$. The left-hand panel shows a disc simulated using standard, Monaghan (1992) SPH with $\alpha\beta$ viscosity, and the right-hand panel shows the *same* initial disc, but simulated using the fully conservative grad-h SPH (Monaghan, 2002; Price & Monaghan, 2004; Springel & Hernquist, 2002). The disc on the left, using standard SPH, has fragmented to form bound objects, while the *same* disc, evolved using grad-h SPH, has not. This could be due to the error in energy conservation that can arise in shock regions in standard SPH (potentially up to $\sim 10\%$ Hernquist 1993), since without the additional grad-h terms in the equations of motion and energy the formalism is not completely conservative. This is interesting, as it *may* suggest that any given value of $\beta_{\rm crit}$ for the fragmentation boundary is very slightly underestimated, but to be sure would require additional simulations, at decreased resolution, with incremental β values along the β fragmentation boundary. An alternative explanation is that it is simply due to some stochastic effect of the simulation, arising, perhaps, from different timestepping calculations.

Within the same SPH implementation, changing how the artificial viscosity is implemented may stop fragmentation for a given value of β, as is the case for the 0.5 million run, at $\beta = 6$, using RTSPH, that fragmented to form one bound clump when using $\alpha\beta$ viscosity, and did not fragment at all when using Monaghan (1997) viscosity. This is shown in Figure 3.13, which shows both of these discs, simulated at a resolution of 0.5 million particles, both with $\beta = 6$ and $\alpha\beta$ viscosity in the left hand panel, and Monaghan (1997) viscosity in the right hand panel.

From this figure, we can see that which viscosity prescription we choose matters when performing convergence testing. Since we are dealing with collapsing

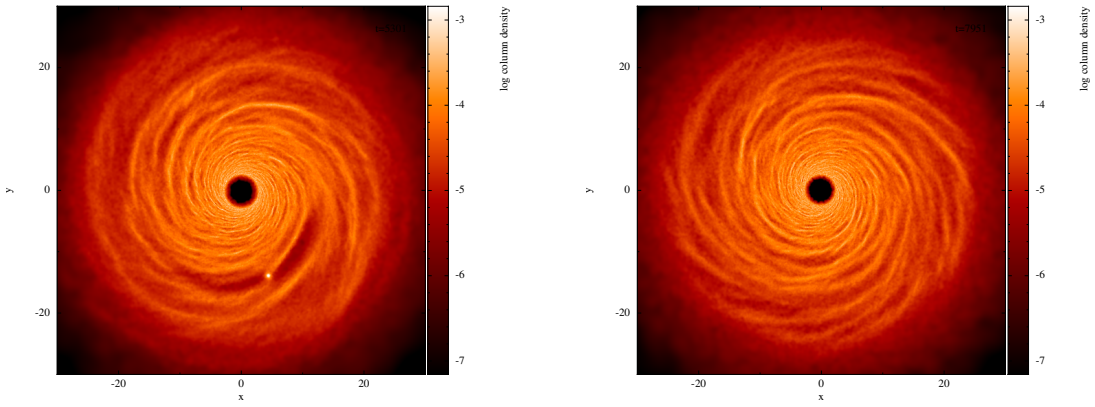


Figure 3.13: Both panels show a 0.5 million disc, simulated with $\beta = 6$, both using the Ritchie & Thomas (2001) SPH formalism. Left hand panel shows traditional $\alpha\beta$ viscosity, right hand panel shows Monaghan (1997) viscosity based on Riemann solvers. We can see that viscosity implementation matters, since the disc on the left fragmented, and the disc on the right hand side did not.

regions, we expect a large density contrast over small scales, and artificial viscosity, in its usual implementation, may have trouble capturing this correctly. Perhaps density is over-estimated which causes accelerated cooling of material onto the collapsing fragment. Monaghan (1997) viscosity is *physically* motivated, with the amount of dissipation calculated by the signal velocity between the two interacting particles. However, standard $\alpha\beta$ viscosity does not have so much a physical motivation as an algebraic one, it "behaves itself", mathematically speaking; it was chosen for its Galilean invariance, conservation of angular momentum and vanishing property for rigid solid-body rotation. It may be the case that, generally speaking, Monaghan (1997) viscosity should be the default viscosity prescription in SPH codes.

More generally, the choice of viscosity prescription is important even when not considering convergence, Figure 3.14 shows the 0.25 million particle disc, evolved

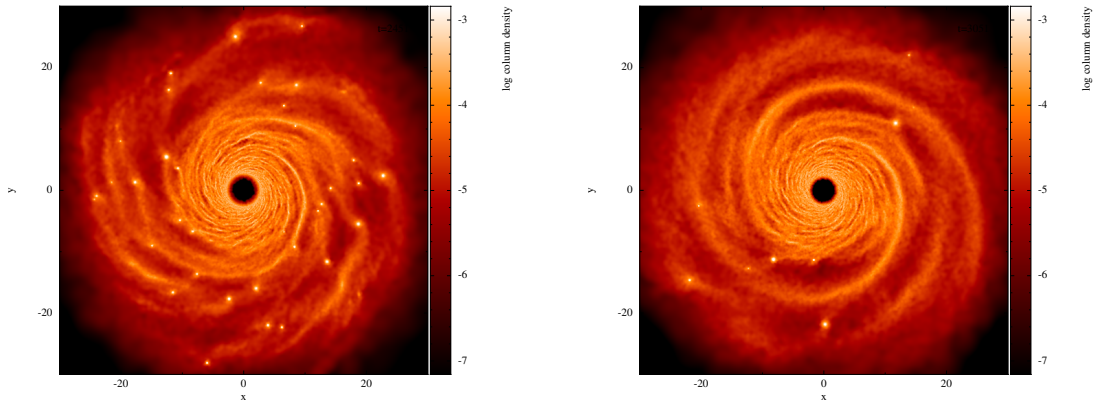


Figure 3.14: Both panels show a simulation with a resolution of 0.25 million particles. Both use Ritchie & Thomas (2001) SPH, and both have $\beta = 3$. Left has $\alpha\beta$ viscosity and right has Monaghan (1997) viscosity. We can see that the disc using viscosity inspired by Riemann solvers (right) has a reduced amount of fragmentation.

using Ritchie & Thomas (2001) and $\beta = 3$, left with $\alpha\beta$ viscosity and right with Monaghan (1997) viscosity. Although both discs have fragmented, they look considerably different, with more fragments in the $\alpha\beta$ version than in the Monaghan (1997) version.

3.11 Discussion and Conclusion

In this chapter, we have discussed the governing principles behind SPH, beginning with the motivation for its invention, its core principles and the discretised equations of hydrodynamics. In our discussion of artificial viscosity, necessary to capture density contrasts correctly, we have discussed a form of viscosity based on Riemann solvers, and how this led to the inception of an artificial conductivity term capable of smoothing discontinuities in thermal energy. We then presented a suite of 102 global SPH simulations of protostellar discs, with and without this artificial conductivity implemented, in an attempt to find a solution to the

convergence problem of $\beta_{\rm crit}$ with increasing resolution. We found that while artificial conductivity *can* prevent fragmentation in discs that without it would fragment, it was not robust, and could just as likely be due to some stochastic effects within the disc.

While pursuing this line of enquiry, we decided to investigate how the implementation of SPH, and artificial viscosity, affected the fragmentation boundary. To this end, we ran a suite of 71 global SPH simulations using three different types of SPH: standard (e.g. Monaghan 1992), grad-h (e.g. Monaghan 2002; Price & Monaghan 2004; Springel & Hernquist 2002) and the Ritchie & Thomas (2001) formalism, which assumes that pressure, rather than density, is constant over the smoothing kernel.

This did not result in a solution to our convergence problem (we discuss the likely solution below, and, in light of this, why our investigation was not successful), however it did highlight a few, important, points about our investigation, namely that different implementations of SPH can give a different answer for the same values (likely due to better or worse energy conservation), and that *how* artificial viscosity is implemented can alter the appearance of the disc, and also reduce the amount of fragmentation present.

Ultimately, it is unsurprising that our varied investigations did not result in convergence. In every simulation, we calculated the cooling rate of each particle by

$$\left(\frac{du_j}{dt}\right)_{\rm cool} = -\frac{u_j}{\tau_c}, \tag{3.106}$$

where

$$\tau_c = \frac{\beta}{\Omega}. \tag{3.107}$$

However, it was realised by Rice et al. (2012b) that using this form of cooling in the energy equation (given by equation 3.27) produces a mismatch in scales. With the addition of equation 3.106, the energy equation now reads

$$\frac{du_a}{dt} = \frac{P_a}{\rho_a^2} \sum_b m_b \mathbf{v}_{ab} \cdot \nabla_a W_{ab} + \frac{1}{2} \sum_b m_b \Pi_{ab} \mathbf{v}_{ab} \cdot \nabla_a W_{ab} - \frac{u_j}{\tau_c}. \tag{3.108}$$

We can see from the final term in this equation that the cooling is not acting on the same length scale as the rest of the equation. In SPH, physical properties such as density and energy should be thought of as smeared distributions around the particle centre. Introducing a term into equation 3.108 that is not convolved with the smoothing kernel means that the equation is inconsistent, and not a true representation of the philosophy of SPH.

In the introduction to this chapter, we described the point at which this work began, and we now describe the point at which this work ended. Rice et al. (2012b) ran a 10 million particle simulation that did not appear to converge to a predicted β value. However, this was an *effective* 10 million particle simulation, done in a ring using 4 million particles. Subsequently, Rice et al. (2014b) ran a full 10 million particle simulation, with the same parameters, and showed convergence to $\beta = 8$ for 10 million particles using their "smoothed cooling" formalism, where the cooling is spread over the particles i in the neighbour sphere of particle j,

such that for particle j we have

$$\left(\frac{du_j}{dt}\right)_{\rm cool} = -\frac{1}{\tau_c}\frac{W(0,h_j)m_j u_j}{\rho_j} \tag{3.109}$$

and for the particles i in the neighbour sphere we have

$$\left(\frac{du_{i\neq j}}{dt}\right)_{\rm cool} = -\frac{1}{\tau_c}\frac{m_i}{\rho_i}W(\mathbf{r}_{ji},h). \tag{3.110}$$

Essentially, in non shock regions, this cooling behaves identically to normal β cooling, but in regions where discontinuities exist, using the interpolated value of the cooling rate will prevent artificial enhancement of cooling in that region (Rice et al., 2012b, 2014b).

Given that in our investigations we did not use this smoothed cooling, it is unsurprising, in light of the convergence achieved by Rice et al. (2014b), that we did not achieve convergence here. The fact that convergence was achieved by Rice et al. (2014b) does, however, mean that our essential understanding of fragmentation is unchanged. The fragmentation boundary is likely to be similar to that found in earlier work (Gammie, 2001; Rice et al., 2005a), implying that gas giant planet formation by fragmentation is unlikely in the inner 50 au of discs (Rafikov, 2005; Rice & Armitage, 2009).

Oh, yes. And things have to be parallel. I see a picture right now that's not parallel, so I'm going to go straighten it. Things must be in order.

Katherine Johnson, NASA Aerospace Technologist

4

Identifying Protostellar Disc Fragments in Smoothed Particle Hydrodynamics Simulations

This chapter contains work that has been submitted for publication in the Monthly Notices of the Royal Astronomical Society. CH is first author, with Duncan Forgan and Ken Rice as second and third co-author respectively.

4.1 Introduction

In this chapter, we analyse a suite of SPH simulations of fragmenting protostellar discs, identifying fragments using two different methods. The first, based on the `CLUMPFIND` algorithm (Smith et al., 2008; Williams et al., 1994), is done using the gravitational potential, and the second is a new method that uses density derivatives. We do not use sinks in our simulations, as by using only SPH particles, we are able to determine the fragment internal structure as it migrates through the disc, and better understand the orbital evolution of the fragment, which is sensitive to its radial mass distribution. We discuss the relative merits of our different detection methods for our simulations, which show a variety of fragmentation scenarios.

4.2 Methods

4.2.1 Smoothed Particle Hydrodynamics

Smoothed Particle Hydrodynamics (SPH) is a Lagrangian hydrodynamics formalism, that evolves a fluid by means of a distribution of pseudo-particles (Gingold & Monaghan, 1977; Lucy, 1977). There are many review articles about SPH (see e.g. Monaghan 1992, 2005; Rosswog 2009), but the basic idea is that each particle has a position, mass, internal energy and velocity, and these parameters can be interpolated over to give fluid quantities at any position. Density is calculated by interpolation over the mass distribution, and pressure is determined using an equation of state with internal energy. Gravitational forces are usually computed using a TREE algorithm (Barnes & Hut, 1989), and then the discretised

energy and momentum equations are solved. Particle velocities are updated using pressure and gravitational forces, and positions are then updated using these velocities. Internal energy changes are computed by calculating $P\mathrm{d}V$ work, viscous dissipation and radiative cooling and heat conduction.

Cooling calculations in SPH are not simple. Accounting for polychromatic radiative transfer within a hydrodynamics simulation is not possible with current computational resources, and even post-processing a single snapshot with radiative transfer is computationally expensive (Stamatellos & Whitworth, 2005b). Historically, approximations to individual features of radiative transfer were used, such as the cooling time formalism: $\dot{u} = -u/t_{\mathrm{cool}}$ (Rice et al., 2003b). Although this parameterisation is useful, allowing us to probe the effects of different cooling timescales in protoplanetary discs, it is somewhat limited, as it only allows us to model energy lost from an SPH particle. Realistically, if energy is lost from one SPH particle, at least some of that energy will be gained by its surrounding neighbours - this is known as radiative transfer.

Since our aim here is to trace the orbital and profile evolution of fragments within protostellar discs, we wish to capture the effects of radiative transfer as far as is feasible. Therefore, the cooling we implement is the hybrid method of Forgan et al. (2009). The details of the algorithm are given in Forgan et al. (2009), however the basic ideas merge the polytropic cooling method of Stamatellos et al. (2007) with the flux-limited diffusion method of Mayer et al. (2007b), which builds on conduction modelling work by Cleary & Monaghan (1999) and the flux-limiter described in Bodenheimer et al. (1990). The biggest advantage is the complementary nature of these two methods: energy loss is handled by

polytropic cooling (which flux-limited diffusion cannot do), and positive energy exchange between neighbouring particles is handled by flux-limited diffusion (which polytropic cooling cannot do). Since each method handles a separate process, there is no "double counting" in any part of the system's overall energy, and these separate parts can simply be summed to calculate the total energy change.

4.2.2 Simulation setup

We run a total of 9 simulations of 0.25 $M_\odot$ discs, with a 1 $M_\odot$ central star, an inner radius of 10 au, an outer radius of 100 au, and a radial density profile of $\Sigma \propto r^{-1}$. All discs are initially identical in global properties, varying only the random number seed used to initialise the disc. All discs are evolved until it is no longer computationally feasible to continue, which in reality means the density of a fragment has become so high that timesteps cannot be computed without switching that mass to a sink particle. However, since here we wish to examine physical and orbital properties of the fragments which are influenced by their radial mass distribution, we do not do this. All of the simulations fragment to form at least two bound objects, and their ultimate configuration is shown in Figure 4.1, which shows 9 column density plots, in physical units, and illustrates a variety of fragmentation scenarios as the simulation's final configuration. We discuss this in detail in section 4.3, but include images now to make the explanation of our methods in section 4.2.3 clearer.

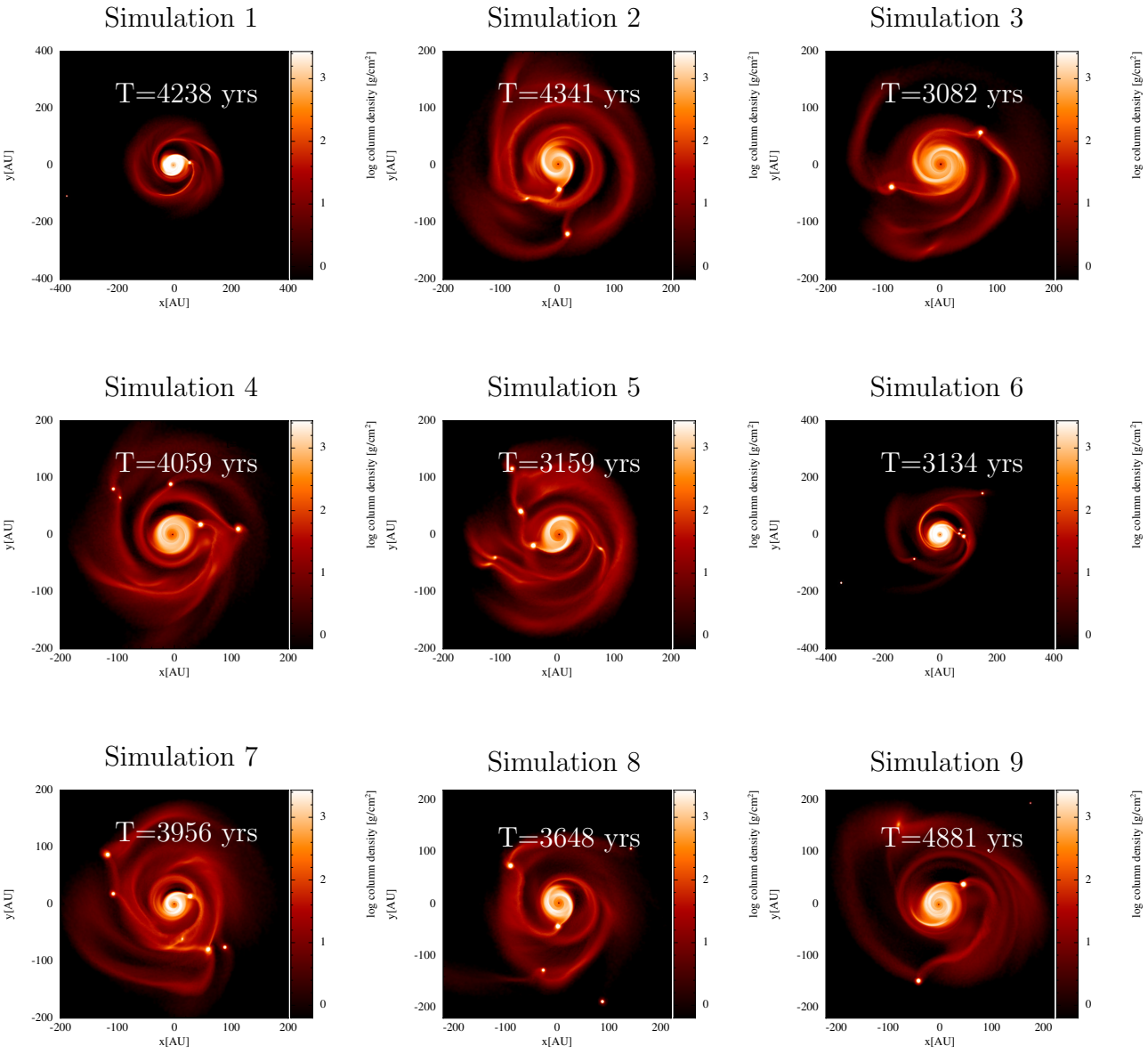


Figure 4.1: Column density plots of the final fragment configuration for all 9 simulations. Despite almost identical initial conditions, there are a variety of ultimate configurations, and a variety of times at which it becomes computationally unfeasible to continue the simulation using only hydrodynamics. Simulations 1 and 6 show an ejected clump at large radial separation, and the top left hand corner of simulation 9 shows a fragment forming just below the threshold detection of our algorithm.

4.2.3 Algorithms

We present here two methods of detecting fragments in SPH simulations, and one method of linking them together between timesteps. Once a fragment has been identified, we then refer to it as a "clump". The first method of detection is based on the clump finding approach of Smith et al. (2008) (which is in turn based on the publicly available `CLUMPFIND` algorithm developed by Williams et al. 1994). The basic idea is to perform an ordered search on SPH particles from high (physical) density to low density. The highest density particle i forms the center of a clump, and if the next particle in the list is a neighbour (i.e. in close spatial proximity), it is also added to this clump. If it is not a neighbour, it forms the center of a new clump. This process is continued to the next most dense particle, until a minimum density threshold is reached.

The search in this manner, from most dense to least dense particle for our SPH simulations of protostellar discs was unsuccessful in identifying clumps in our simulation. We are faced with a different scenario to Smith et al. (2008), who used their algorithm in molecular cloud cores. Once our discs have evolved enough to fragment, the inner disc is sufficiently dense to start being identified by `CLUMPFIND` as the beginning of a new clump. Since the algorithm is based on a friends-of-friends approach, this resulted in spurious clump detection throughout the disc.

This problem was solved, to some extent, by using the gravitational potential of the particles, rather than the density, for the ordered search. We discuss this method in section 4.2.3.1. The inability of this method to identify low mass, fluffy fragments, or fragments that are so deep in the potential well of the central

star that they are ultimately tidally destroyed, prompted the development of an approach that could correctly identify such fragments. The approach uses a gridded derivative search of the SPH interpolated density of the particles, and is discussed in section 4.2.3.2.

Finally, in both cases, the clumps are linked between timesteps using a merger tree algorithm, typical of dark matter halo tracers in cosmological simulations (see, for example, Srisawat et al. 2013). This process is detailed in section 4.2.3.3.

4.2.3.1 Gravitational potential search (`CLUMPFIND`)

Broadly speaking, clumps are created with a unique integer identifier (ID) at the local minima of the gravitational potential, so long as there are at least a minimum number of neighbour particles above some defined "noise" level. For our purposes, we define this critical number as $n_{\rm critical} = n_{\rm mean} - 5n_{\sigma}$, where $n_{\rm mean}$ is the mean number of neighbours each particle has, and n_{σ} is the standard deviation of the number of neighbour particles. We do this because the neighbour lists of particles at low density can be sparse due to the algorithms used to calculate the smoothing length, h.

We begin by creating a clump at the location of the central sink particle (star). All particles are then sorted by their gravitational potential energy, and we loop over the particles in order of most negative to least negative gravitational potential energy (i.e. most bound to least bound). We select the particle, i, with the most negative potential energy, and as long as it does not already belong to a clump (in which case, we exit and select the next particle), we iterate over particle i's neighbours, j. If the majority of the neighbours j (>50%) are in a clump k, the

particle is also in clump k. We assign the particle ID of i to $\mathrm{ID}_i = k$ and exit. If the majority of i's neighbours are not in a clump, then since i is the most bound particle it starts a new clump l, provided that $n_{\mathrm{neighbours}} > n_{\mathrm{critical}}$. We then loop over particle i's neighbours j, assigning $\mathrm{ID}_j = l$ so long as j is not already in a clump. We then proceed to next most bound particle, $i + 1$, and repeat (Forgan et al., 2017).

4.2.3.2 Density Derivative Search

In this method, we compute a 2D grid in cylindrical r and ϕ coordinates, and bin all particles into these grid cells. The maximum density of a particle in each cell is then taken to be the peak density in that cell, ρ_i, which gives us a 2D sheet describing density maxima. The number of bins required to reasonably identify all clumps varies due to the stochastic nature of the simulations, more fragments with low density at ~ 100 au require a larger number of bins to properly resolve, with our resolution criterion being that the number of clumps ultimately detected by the search is equal to the number of clumps that are determined "by-eye". If fewer clumps are detected by the search than "by-eye", then the resolution is increased until we detect these low-mass, low-density clumps. A typical resolution is 10,000 radial bins and 7200 azimuthal bins.

We next take the derivative of the peak density in that cell with respect to r and ϕ. As this is noisy in density space, we smooth these derivatives, equivalent to making a new signal, where the element is now the average of n adjacent elements,

such that:

$$\frac{\partial \rho_i}{\partial r} = \frac{1}{n} \sum_{j=-\frac{n-1}{2}}^{j=\frac{n-1}{2}} \frac{\partial \rho_{i+j}}{\partial r} \tag{4.1}$$

and

$$\frac{\partial \rho_i}{\partial \phi} = \frac{1}{n} \sum_{j=-\frac{n-1}{2}}^{j=\frac{n-1}{2}} \frac{\partial \rho_{i+j}}{\partial \phi} \tag{4.2}$$

Despite its simplicity, this approach is best at removing white noise while keeping the sharpest step response. The value of n (~ 100 is typically sufficient), like the number of bins in which to bin the data, must be optimised by the user to get the best compromise between smooth data, which removes false peaks, and data which is sensitive enough to identify small clumps.

We then use this smoothed derivative to identify clumps, which will be a "real" peak in the density. A real peak is identified by a sustained zero-crossing of the derivative with a negative gradient. Peaks due to noise will also have zero-crossings, but they are sustained for fewer bins than real peaks. These false peaks can be eliminated by requiring that the zero-crossing is sustained for m bins, with m optimised by the user to remove most (if not all) false detections while still detecting less dense clumps.

The radial search is shown in Figure 4.2, which shows the radial density profile (blue solid line) and the derivative of the radial density profile (dashed red line) for the disc shown in simulation 6 of Figure 4.1. The zoomed section shows the peak of a clump, with the negative gradient zero-crossing of the derivative identifying the peak. Once the particle marking the center of a clump (i) has been identified, we add all of i's neighbour particles j to that clump. We now loop

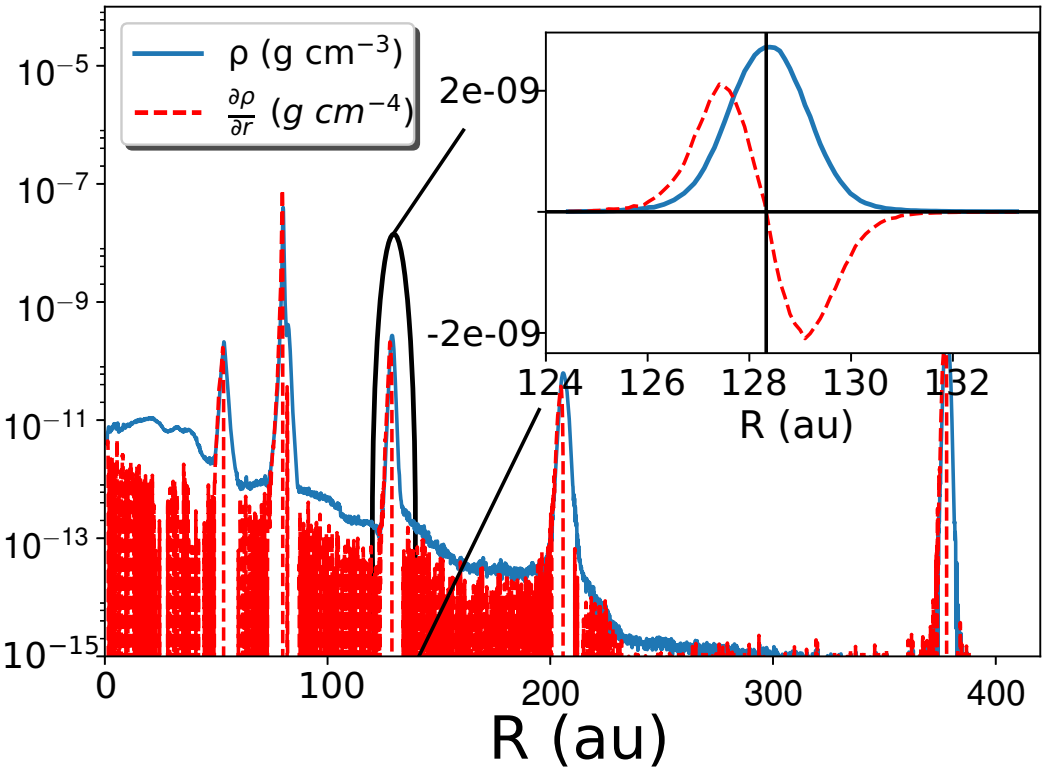


Figure 4.2: Our density derivative search method on the disc shown in Figure 4.1, simulation 6. The solid blue line shows the radial density profile of the disc, and the dashed red line shows the derivative of this with respect to r. The zoomed region demonstrates how the negative zero crossing of the derivative identifies one of the real density peaks.

over all particles which form that clump, adding their neighbours to this clump as well. We repeat this until we reach some density threshold. We found that adding neighbour particles until more than half the particles in the neighbour sphere are less dense than the inner 1 au of the disc produced good results. Once we have identified the bulk of the clump, we then proceed with a potential search described in section 4.2.3.1, which determines to which clumps the rest of the unidentified particles in the simulation belong.

4.2.3.3 Merger Tree

At this point, we have a set of clumps in each timestep, and we need to track them over the duration of the simulation. To do this, we use a standard algorithm from halo tracking in cosmological simulations (see, e.g., Springel et al. 2001). Each clump, in each timestep, is given an integer ID by our algorithm. So that we can trace the evolution of this clump throughout the simulation, these IDs must be linked. Since we are modelling fluid through the use of pseudo-particles, the particles, that make any given clump, change between timestep file-dumps, sometimes substantially. To link clumps, the crucial factors are the most-bound particle MBP, and shared member fraction (SMF). In our density-derivative search, we actually trace the most-dense particle, rather than the most-bound particle, but we use the term interchangeably to avoid the introduction of unnecessary acronyms. To be identified as the same clump between timesteps, they must share the MBP and have an SMF of $>50\%$. In some particularly volatile simulations, when using the density derivative method, the MBP may change, and the SMF may be $<50\%$. In this case, some of the clumps need to be manually linked during post processing by tracing a group of particles in each

clump in each timestep. The basic algorithm is as follows:

1. Loop over clumps i in previous timestep lastdump.
2. Find clump j, in this timestep thisdump, which contains the MBP from clump i in lastdump.
   (a) If MBP_i does not belong to any clump in thisdump, clump i is not present in thisdump.
   (b) If MBP_i belongs to clump j in thisdump, and clump i and j share at least 50% of particles, then $\text{ID}_j = \text{ID}_i$, and the clumps are linked between the two timesteps.
   (c) If MBP_i belongs to clump j in thisdump, but clumps i and j do not have SMF >50% of particles, then clump j keeps its present ID.
3. End loop over previous timestep.
4. Loop over clumps in thisdump, checking for clumps with no progenitors in lastdump. Increment the maximum number of distinct IDs by the number of new clumps, and assign each clump the correct ID.

4.3 Results

We identified clumps in our simulation using two different search methods, an ordered potential energy search (OPS) based on Smith et al. (2008), and a novel approach based on a 2D density derivative search (DDS). Our first conclusion is that searching using the ordered potential of the particles only detects clumps that survive the duration of the simulation. However, using the DDS method,

destroyed clumps are also detected. This is shown in Figure 4.3, which shows the total change in semi major axis, from when the clump is first detected to when the clump is last detected (or destroyed), against the total time for which the clump exists, i.e. from initial identification to the end of the simulation, or the last timestep in which it is identified, if it is destroyed. Larger markers indicate more massive clumps. The left hand panel shows the DDS results, the right hand panel shows the OPS results. Circular markers indicate clumps that have survived until the end of the simulation, square markers indicate clumps that are destroyed, and triangle markers indicate a clump that merged into another clump. There are no identical markers in both plots because the different algorithms detect the clumps at different times, therefore they migrate different distances.

In addition to containing no destroyed clumps, the OPS sample also has a relative insensitivity to clumps which will have a final mass of less than ~ 5 M_J, and detects most of the clumps later in the simulation. This is shown in Figure 4.4, which shows mass accretion histories for clumps in each of the 9 simulations, as detected by the OPS. By comparison, Figure 4.5 shows the same 9 simulations, but with the mass accretion histories of the clumps determined using the DDS method. As can be seen, many low-mass clumps evade detection entirely under the OPS method; for example, simulation 6 has an additional 3 clumps that are not detected by the OPS, and those that are detected are generally detected later, such as clump 2 in simulation 2, which is detected ~ 400 years later in the OPS than in the DDS.

This is due to the nature of the potential search. Figure 4.6 shows the radial gravitational potential energy profile of the disc in Figure 4.1, simulation 6. Since

the OPS proceeds from the particle with the most negative gravitational energy to the most positive, the OPS detects the clump at ~ 80 au in Figure 4.6 first, and then detects the clump at ~ 375 au second. However, it fails to detect the clumps at ~ 50 au, ~ 125 au and ~ 200 au. This is because particles at the potential energy of the main body of the disc are identified as belonging to the clumps with the most negative potential energy during the neighbour check described in section 4.2.3.1.

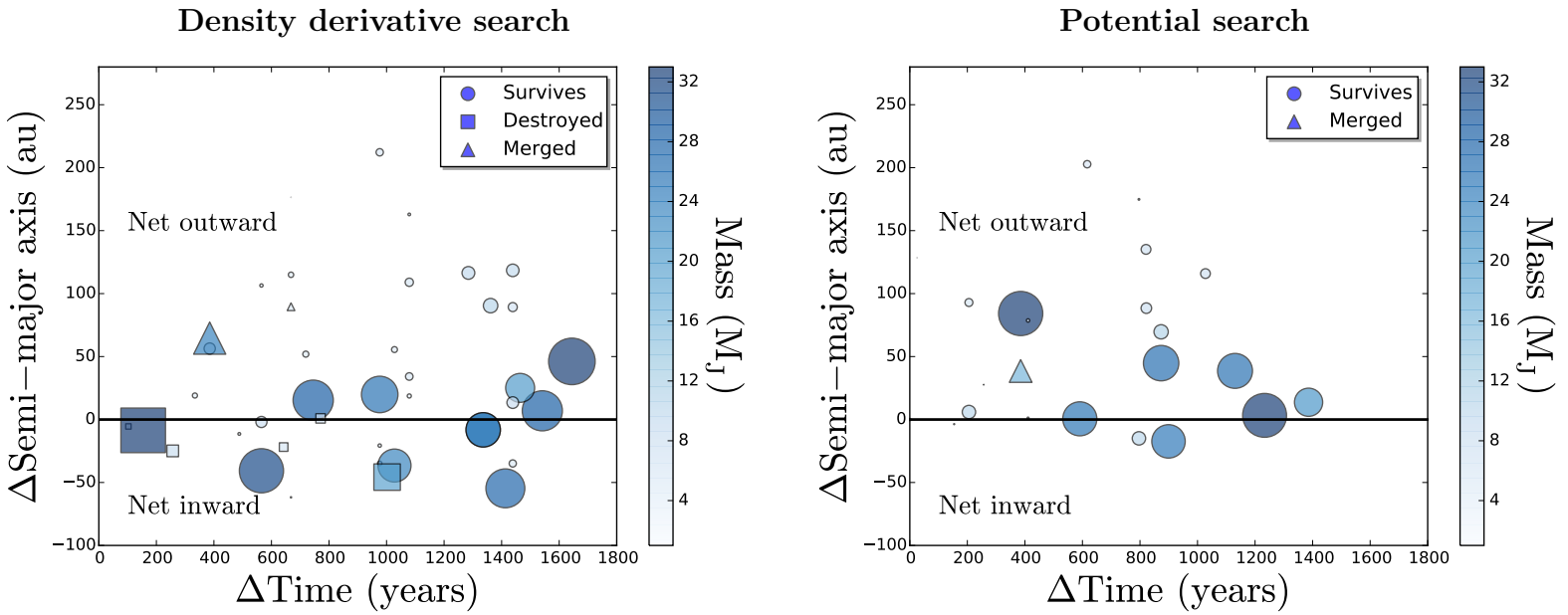


Figure 4.3: Plots showing total change in semi-major axis, for all clumps in all 9 simulations, plotted against the time between initial identification, and either the end of the simulation, or the last timestep in which they are identified, if they are destroyed or merged. Larger markers correspond to more massive clumps. The left hand panel shows clumps as they are identified by our density derivative search method, and the right hand panel shows clumps as they are identified using our ordered potential search method. Circular markers indicate the clump survived the duration of the simulation, square markers indicate tidally destroyed clumps, and triangle markers indicate the clump was subsumed by another clump. The right hand panel shows that clumps that are ultimately destroyed are not detected by the ordered potential search. The left hand panel shows that $\sim 20\%$ of fragments are ultimately tidally destroyed. There are no identical markers in both plots because the clumps are detected at different times in the simulation, and thus migrate different distances.

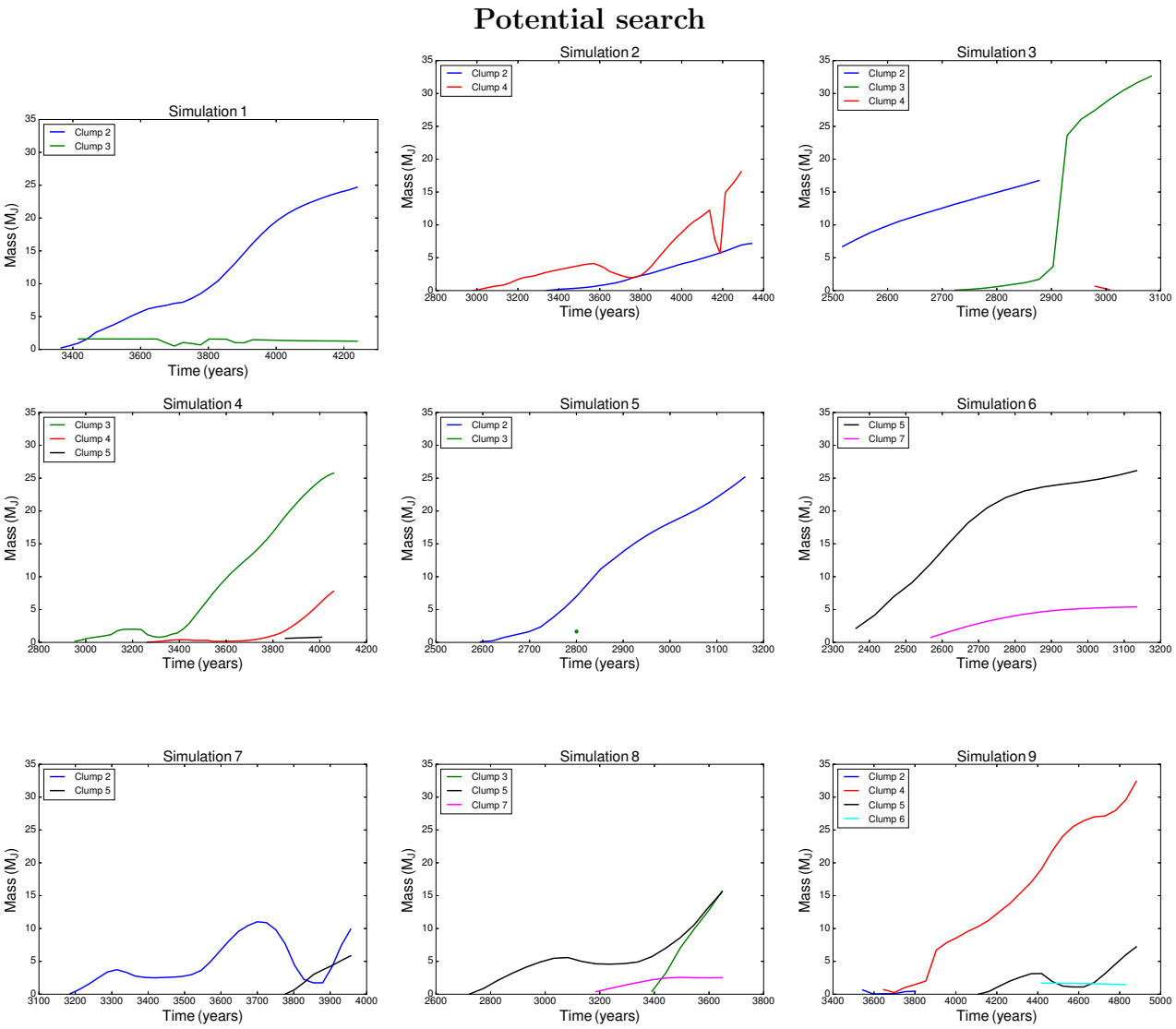


Figure 4.4: Mass accretion history for all clumps in all simulations, using only the ordered gravitational potential energy search. Fewer clumps are detected by this method than by using the density derivative, but those that are detected are likely to survive for a long time. Clumps are generally detected later in their evolution using this method, when their gravitational potential energy is negative enough to have neighbour particles assigned to them before they are assigned to the main body of the disc.

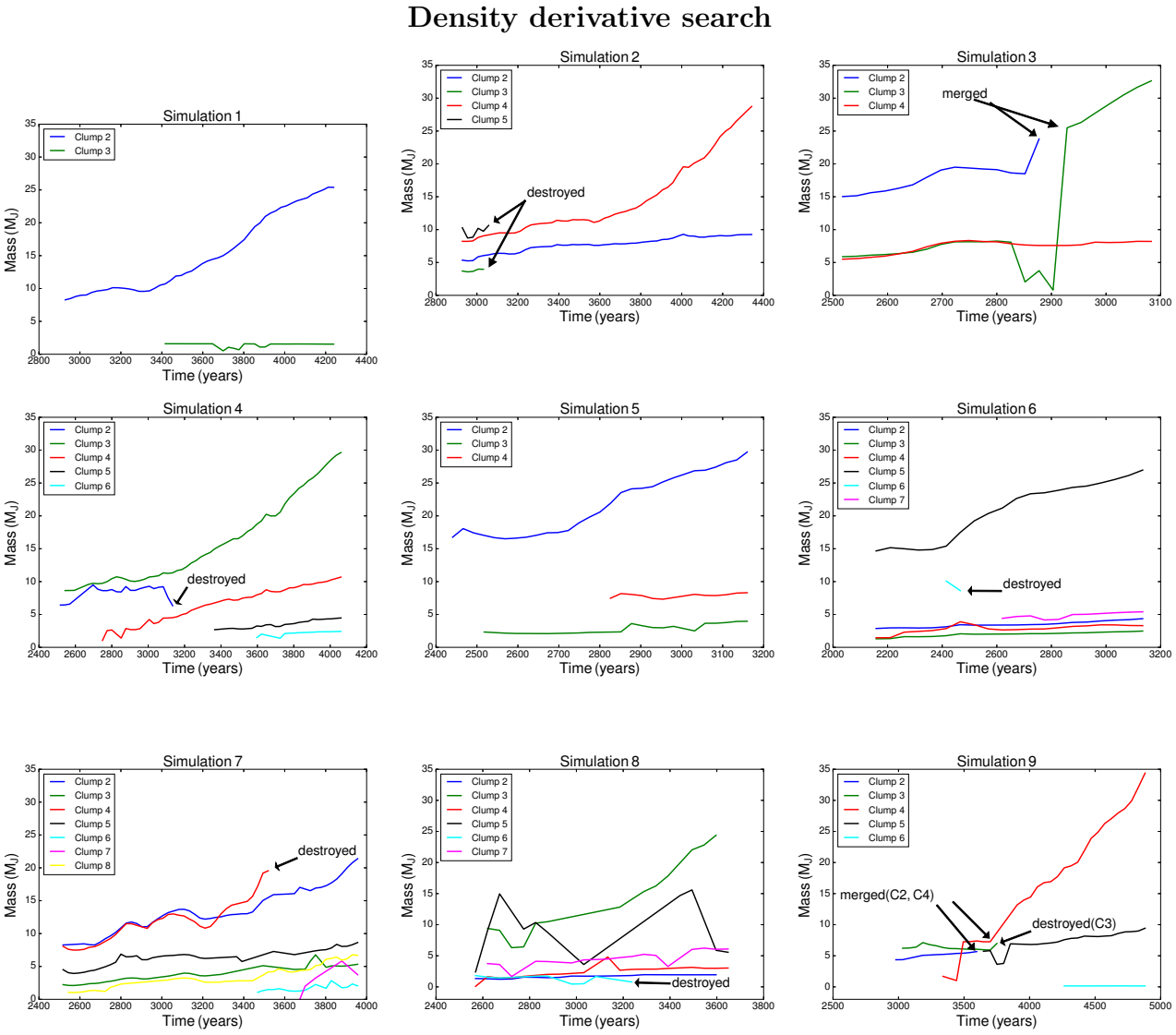


Figure 4.5: Mass accretion history for all clumps in all simulations, found using the density derivative search method. More clumps are detected this way than by using the ordered potential search, as clumps buried in the potential well of the disc are identified early by their density peak. As can be seen by comparison with Figure 4.4, this search method is sensitive to low mass clumps, is sensitive to all clumps earlier in their evolution, and, by comparison with Figure 4.3, we can see this method is also able to detect clumps which are ultimately tidally destroyed.

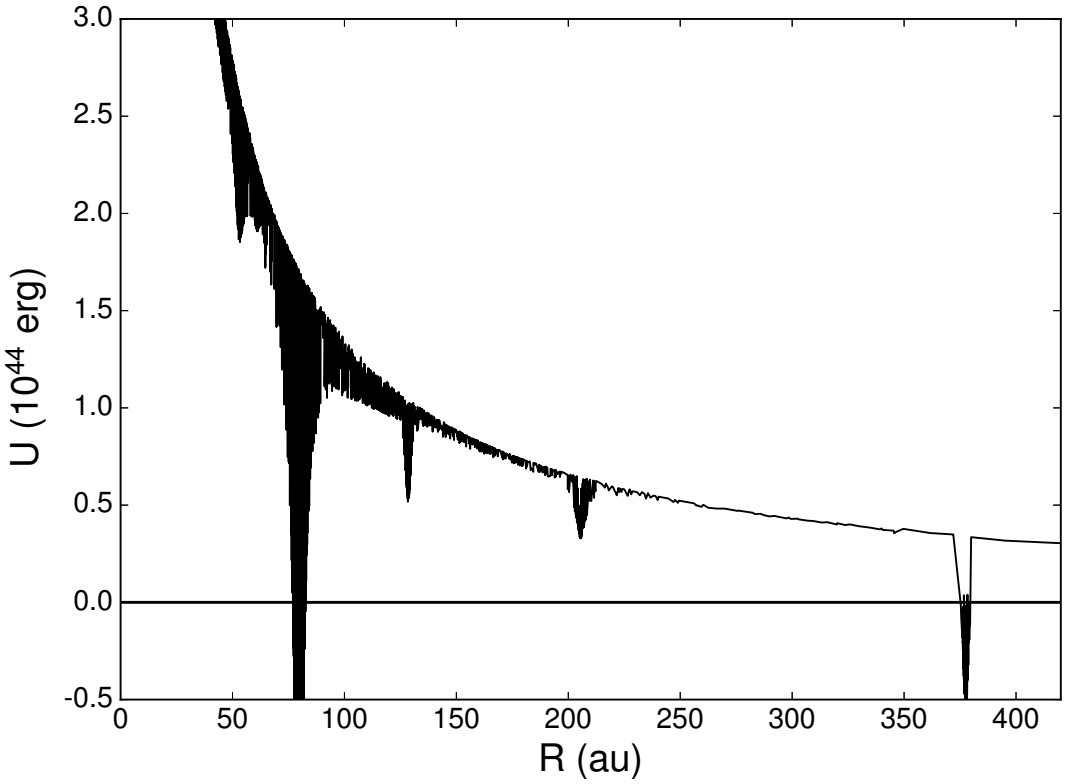


Figure 4.6: Radial profile of gravitational potential energy for the disc shown in Figure 4.1, simulation 6. Only two clumps, at $r \sim 80$ au and $r \sim 375$ au have a sufficiently deep potential well to be identified by our ordered potential search. The other three clumps, at $r \sim 50$ au, $r \sim 125$ au and $r \sim 210$ au are identified as belonging to the main body of the disc. Their detection in the density derivative search, but not in the ordered potential search, indicates that they would likely be tidally destroyed.

Then, when the particles belonging to clumps at 50, 125 and 200 au are checked, they are found to already belong to either the main body of the disc (i.e. the central sink) or one of the clumps with the deepest potential well.

Although this could be fixed by adopting a gridded approach to the potential search (thereby eliminating the dominating effect from the clumps with the largest potential well), OPS has a desirable feature, namely demarcating clumps that are likely to survive the simulation, and those that are not.

The OPS method's insensitivity to small clumps, and reliance on deep potential wells for identifying the body of the clump, mean that fewer clumps are detected by the OPS method, and those that are identified are often initially identified at artificially small masses ($\sim 10^{-3}$ M_J), as there is only a small amount of mass with a potential well deep enough to be identified. This can be seen in Figure 4.4, which shows mass accretion histories for all 9 simulations using the OPS method. In every simulation, what ultimately grows to be the largest clump is initially identified with a mass well below the Jeans mass. The DDS search method, however, does a better job of correctly identifying the mass associated with the young clumps, typically identifying between 5 M_J and 10 M_J of mass. This is shown in Figure 4.5, which shows the mass accretion histories for all 9 simulations, as identified using the DDS method.

In addition to the difference in measured initial clump mass, both methods differ in the final mass attributed to clumps. Typically, the growth is smoother for the OPS method, since once mass is deep in the potential well it is unlikely to change. However, if the gravitational potential energy profile of the disc changes then so too will the attributed mass of the clumps.

Another feature of the DDS method is its ability to identify low mass, low density clumps that do not have a strong signal in their potential. This can be seen in Figure 4.5, simulation 6, which identifies an additional three clumps compared to Figure 4.4, simulation 6. Comparing Figure 4.6 and Figure 4.2, we can see that these clumps have much stronger signals in their radial density profiles compared to their radial potential energy profiles. This is a useful predictive feature, since the left hand panel of Figure 4.3 and all of Figure 4.5 show that $\sim 20\%$ of the

fragments in our simulations are tidally destroyed (we discuss the implications of this for population synthesis in section 5.2.1), none of which are detected in the OPS method.

Therefore, if a clump is detected in the DDS and not in the OPS, it is indicative that either the clump will stay relatively low mass and not accrete further, or that it will be tidally destroyed.

4.4 Discussion and Conclusion

We have presented one original method for identifying fragments in a simulation, the density derivative search, and one method adapted from the `CLUMPFIND` (Forgan et al., 2017; Smith et al., 2008; Williams et al., 1994) algorithm. We ran 9 SPH simulations of a 0.25 $M_\odot$ disc around a 1 $M_\odot$ star, each with a surface density profile of $\Sigma \propto r^{-1}$, an inner radius of 10au and an outer radius of 100 au. Each simulation was run for as long as computationally feasible without converting dense regions to sink particles, since we wished to calculate orbital properties for our fragments, which are sensitive to their radial mass distribution.

Each disc fragmented to form at least 2 bound objects, and we analysed the fragments (which we call clumps, once they have been detected) using the density derivative search and the adapted `CLUMPFIND` method. We have shown that these two methods are complementary, as the density derivative search is able to detect low mass clumps, and clumps that are ultimately tidally destroyed, while the search using the adapted `CLUMPFIND` method filters out clumps which are unlikely to survive the simulation, but also has a relative insensitivity to low mass clumps.

4.5 Acknowledgements

We would like to thank Daniel Price for his publicly available SPH plotting code `SPLASH` (Price, 2007), which we used to produce Figure 4.1. KR gratefully acknowledges support from STFC grant ST/M001229/1. DF gratefully acknowledges support from the ECOGAL project, grant agreement 291227, funded by the European Research Council under ERC-2011-ADG. The research leading to these results also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 313014 (ETAEARTH).

Truth suffers from too much analysis.

Frank Herbert, Dune Messiah

5

Analysing Protostellar Disc Fragments in Smoothed Particle Hydrodynamics Simulations

In this chapter, we build on the work of Chapter 4, where we described two methods for identifying fragments in SPH simulations. Now that we have outlined the details of our detection methods, we proceed to characterise our simulations. We compare our simulations to population synthesis models, and outline the orbital and spin properties of our simulations. We describe the density and

temperature profiles of our fragments, and draw conclusions about what this means for current population synthesis models. This chapter contains work that has been submitted for publication in the Monthly Notices of the Royal Astronomical Society. CH is first author, with Duncan Forgan and Ken Rice as second and third co-author respectively.

5.1 Introduction

There are two distinct modes of planet formation in protostellar discs. The first, and most widely accepted, is the core accretion model (CA) (Hubickyj et al., 2005; Pollack et al., 1996). In this model, growth begins with dust grains of ~ 1 μm that coagulate rapidly into larger particles, ultimately settling into the disc midplane where there is enough material for them to grow to kilometre-sized planetesimals. These planetesimals can then grow via collisions into planetary cores, and if sufficiently massive, and if the gas disc has not dissipated, will accrete a gaseous envelope, ultimately becoming a gas giant planet (Lissauer, 1993; Pollack et al., 1996).

Most observational evidence favours this formation mechanism. For example, gas giant planets are preferentially found around metal-rich stars (Santos et al., 2004), with an empirical relationship that quantifies the probability, $\mathcal{P}$, of gas giant planet formation as

$$\mathcal{P} = 0.03 \times 10^{2.0[\mathrm{Fe/H}]}, \tag{5.1}$$

where [Fe/H] is the metallicity of the host star relative to solar metallicity (Fischer

& Valenti, 2005).

Numerical work (Cai et al., 2005) has suggested that this would not be the case if the second mode of planet formation, gravitational instability (GI), were the dominant formation mechanism of these planets, since an increase in metallicity responds to a decrease in cooling rate, resulting in weaker GI activity. This ultimately decreases the likelihood of these systems fragmenting, since weak GI corresponds to smaller stresses in the disc. On the other hand, it has also been shown that metallicity variation makes very little difference to the occurrence of fragmentation (Boss, 2002). However, numerical work by Mayer et al. (2007a) has shown that there is a dependency on molecular weight for disc fragmentation, and therefore fragmentation should happen more favourably around metal-rich stars. So, overall, the picture is far from clear.

In the GI scenario, gas giant planets and brown dwarfs form by direct gravitational collapse in the gaseous protostellar disc (Boss, 1997, 1998; Cameron, 1978; Kuiper, 1951). This happens rapidly, in a relatively early phase of the disc's life, when it is massive enough to be self-gravitating. The advantage of this mechanism is its rapidity; gas giants are able to form on timescales shorter than typical disc dispersion timescales ($\sim$ 5 Myr; Haisch et al. 2001b). While CA is certainly the most accepted model, there are barriers to grain growth at several length scales which seem to indicate difficulty in forming planetary mass objects within the disc lifetime. The most famous of these is the so-called *metre barrier*; as grains increase in size, so do their relative velocities, which makes fragmentation, rather than coagulation, the most likely outcome. A promising solution to this problem is the *pebble accretion theory* (Lambrechts & Johansen, 2012; Levison et al.,

2015). Pebbles are grouped together due to the *streaming instability* (Youdin & Goodman, 2005), whereby pebbles orbit at Keplerian velocity, but the gas is pressure supported from the host stellar radiation, causing the gas to orbit at sub-Keplerian speeds. Feeling a headwind, pebbles slow, and as more pebbles migrate inwards, they group together, as the headwind reduces locally. When sufficiently large, these groups of pebbles gravitationally collapse (Youdin, 2011), and can begin to accrete pebbles until they form giant planet cores (Lambrechts & Johansen, 2012).

At smaller scales, the *bouncing barrier* prevents coagulation of dust grains, as particles of a given size, above a certain velocity, are more likely to bounce off each other than they are to coagulate. This results in growth typically halting at around $\sim$ 1 mm in size. However, there is evidence to suggest that this could be beneficial to planetesimal formation, since the introduction of a few $\sim$ cm sized grains (e.g, through radial drift) can act as catalyst to grain growth, sweeping up grains, while preventing the growth of too many larger objects which would otherwise smash each other apart (Windmark et al., 2012).

It is generally accepted that disc fragmentation is very unlikely in the inner regions (<50 au) of a protostellar disc (Rafikov, 2005). However, the outer regions of protostellar discs may well be susceptible to fragmentation, offering a formation mechanism for directly imaged planets such as those in the HR8799 system (Kratter et al., 2010b; Marois et al., 2008; Nero & Bjorkman, 2009). Core accretion models struggle to explain objects such as those in HR8799, with four planets orbiting at 14, 24, 34 and 68 au, with masses of $\sim$ 5 M_J (Marois et al., 2008, 2010), since there is not thought to be enough material to form such massive

objects at these distances. Additionally, the growth timescales of such objects, through core accretion, typically exceed disc lifetimes by a factor of at least ~ 3, using conservative estimates (Pollack et al., 1996). Gravitational instability may, perhaps, offer an explanation as to the formation mechanism of these systems.

However, it has been suggested that disc fragmentation rarely forms planetary mass objects (Rice et al., 2015), with some hydrodynamics simulations (Stamatellos & Whitworth, 2009) suggesting objects formed by this mechanism quickly grow to brown dwarf masses ($M > 13$ $\mathrm{M_{Jup}}$), with lower limits placed on the fragment mass of $\sim 3 - 5$ $\mathrm{M_J}$ (Forgan & Rice, 2011; Kratter et al., 2010b). This is compounded by the recent possible observation, for the first time, of disc fragmentation in action (Tobin et al., 2016), which shows the birth of three protostars that are well above the upper limit of the planetary mass regime.

The recent reformulation of the GI scenario in to what is now known as "tidal downsizing" (Boley et al. 2010, 2011; Nayakshin 2010a, 2011a,b) does, however, have positive implications for producing low-mass planets at low semi-major axes. The key is to consider the subsequent evolution of fragments into planetary embryos, through dust growth, radial migration and tidal disruption. Forgan & Rice (2013b) combined the physical processes of tidal downsizing with semi-analytic models of disc evolution (Rice & Armitage, 2009) to produce the first population synthesis model for planets formed through GI. Given the similarities between these fragments and "first cores" (see, e.g. Masunaga et al. 1998), they were modelled as polytropic spheres, with polytropic index $n = 1.5$.

They found that $\sim 40\%$ of fragments that formed are ultimately tidally destroyed

by the central star, and of those that survive $\sim 40\%$ are gas-giant planets with solid cores of $5 \sim 10$ earth masses, and the rest are brown dwarfs with no solid core. They also found that low mass embryos tend to remain at larger semi-major axes due to the tidal downsizing process. Out of over 1 million fragments, there was only one terrestrial type planet (core with no gaseous envelope). These results are inconsistent with GI being the dominant planet formation mechanism, but they are certainly consistent with GI forming brown dwarfs and gas giant planets at large radii.

Population synthesis models, by necessity, make simplifying assumptions about the physics that governs the evolution of each planetary system. In particular, interactions between forming planetesimals, and the interaction of the disc with these planetesimals, are not included in the population synthesis models of Forgan & Rice (2013b) that we discuss here. How important these interactions are in determining the final orbital configuration of a system is something that should be carefully considered before further developments are made to such a model. Quantifying the importance of these interactions is difficult, but some headway can be made by performing SPH simulations of fragmenting protostellar discs, and carefully tracking the evolution of fragment orbital and physical properties throughout the duration of the simulation.

The chapter is ordered as follows: We compare our results to current gravitational instability population synthesis models in section 5.2.1. We outline orbital and spin properties of our fragments in section 5.2.2 and 5.2.3 respectively. We describe density and temperature profiles of fragments with the most interesting histories in section 5.2.4, and discuss our findings and conclude in section 5.3.

5.2 Results

We ran a total of 9 SPH simulations with almost identical initial conditions, differing only in the random number seed used to initialise the disc. Column density plots of the 9 simulations are shown in Figure 4.1 on page 157. Despite the almost identical initial conditions, there is a large variation in final configurations and number of clumps in the system.

Our results sections fall into two broad categories- we discuss the implications of our results for current population synthesis models, comparing our clump mass and semi-major axis functions to existing population synthesis models. We show the clump interaction needs to be included in GI population synthesis models as early as during the gas phase, as scattering plays an important role.

Finally, we discuss interesting events in the simulations themselves. We introduce a piece of nomenclature now, to avoid confusion, that SaCb means simulation a, clump b. This abbreviation is given in the title of any plot of a specific clump. Note that our clump numbering begins at 2, since clump 1 is the star+disc system. We also state now, for clarity, that any mass stated for our clumps *includes* unbound material. This is deliberate, in order to track more of the mass of the clump. Furthermore, what is currently unbound material around the clump, at these very early times, may eventually lose spin angular momentum through interactions with material in the disc, ultimately becoming part of the clump. By including the unbound material, we trace more of this process from earlier times. For high mass clumps (> 20 M_J), the amount of unbound material is small, typically around 10% or so. This is larger for lower mass clumps, up to

$\sim 25\%$ of material identified may be unbound, rising to 40% in clumps ocurring in particularly volatile simulations that have many clumps, as their formation is often disrupted by interactions with other.

We look at the orbital properties of the clumps, and discuss clump mergers and tidal destructions by the central star. We show that: (1) destruction and merging are fairly common, (2) interactions between clumps can result in a clump changing its direction of spin from prograde rotation to retrograde rotation, and (3) retro-rotating clumps typically have more dramatic changes in their radial temperature profiles than prograde clumps.

5.2.1 Comparison to gravitational instability population synthesis models

We ran the Forgan & Rice (2013b) GI population synthesis (GIPS) models for 4000 years, which is comparable to the timescale for which our SPH simulations are able to run until no longer computationally feasible. The opacity has the form (Masunaga & Inutsuka, 1999; Nayakshin, 2010b)

$$\kappa(T) = \kappa_0 \left(\frac{T}{10\,\mathrm{K}} \right)^{p_k}, \tag{5.2}$$

where κ_0 is the opacity at $T = 10$ K, and is typically 0.01 cm^2 g^{-1} for nearby molecular clouds (Masunaga & Inutsuka, 1999). For interstellar dust grains, $1 < p_k < 2$ is typical, and we set $p_k = 1$. In addition, we impose that the disc is not truncated after fragmentation, one of the options examined in Forgan & Rice (2013b).

The full details of the model are described in Forgan & Rice (2013b), but we briefly outline the general method here.

1. Run a series of 100 au self-gravitating disc models (in this case, 100 were run). These are the 1D models described in Rice & Armitage (2009), that evolve viscously according to the equations in Lynden-Bell & Pringle (1974); Pringle (1981), but with the addition of a photoevaporative wind term, $\dot{\Sigma}_{\rm wind}$, so the viscous evolution is now described by

$$\frac{\partial \Sigma}{\partial t} = \frac{3}{r}\frac{\partial}{\partial r}\left[r^{\frac{1}{2}}\frac{\partial}{\partial r}(\nu \Sigma r^{\frac{1}{2}})\right] - \dot{\Sigma}_{\rm wind}, \tag{5.3}$$

   where in a steady-state, this can be integrated to obtain the accretion rate

$$\dot{M} = 3\pi\nu\Sigma. \tag{5.4}$$

2. Find the minimum radius at which fragmentation will occur using the Jeans criterion (Forgan & Rice, 2011), where the Jeans mass, $M_{\rm J}$, is given by

$$M_J = \frac{4\sqrt{2}\pi^2}{3{\rm G}^2}\frac{Q^{\frac{1}{2}}c_s^4}{\Sigma\left(1+\frac{1}{\sqrt{\beta_c}}\right)}, \tag{5.5}$$

   and create fragments with masses equal to $M_{\rm J}$. These fragments are placed at separations of a few Hill radii (with the exact spacing randomly sampled).

3. Evolve the entire system, including effects of migration, tidal disruption, grain sedimentation, core formation, turbulence and fragment contraction, according to the equations that outline the "Tidal Downsizing" scenario

outlined in Forgan & Rice (2013b); Nayakshin (2010a, 2011a). Although the fragments may be evolved in parallel, they are evolved *separately*, a major limitation of this model which we discuss extensively in the conclusion to this Chapter.

Figure 5.1 shows initial and final mass and semi-major axis distributions for the GI population synthesis model and for the clumps in our SPH simulation. The red, shaded histogram is the population synthesis, blue outline is the SPH clumps. The left hand panel of Figure 5.1 shows the initial and final semi-major axis distributions for both samples, and the right hand panel shows initial and final mass distribution for both samples. We cut off the tail of the initial and final masses beyond > 35 M_J, since we cannot feasibly simulate masses above this without switching to sink particles.

Comparing the initial mass and semi-major axis distributions in the GIPS models to our SPH clumps, it would initially appear that clumps in SPH simulations form much further out, and at much lower masses, than in the population synthesis models. In reality, this is somewhat a limitation of our identification algorithms, as, at very early times, the clumps can escape detection because of their low density/less negative gravitational potential energy, such that they have already undergone some radial migration before they are detected. If they have undergone sufficient radial migration, they may be far enough out in the disc to not accrete much material, hence remaining low mass.

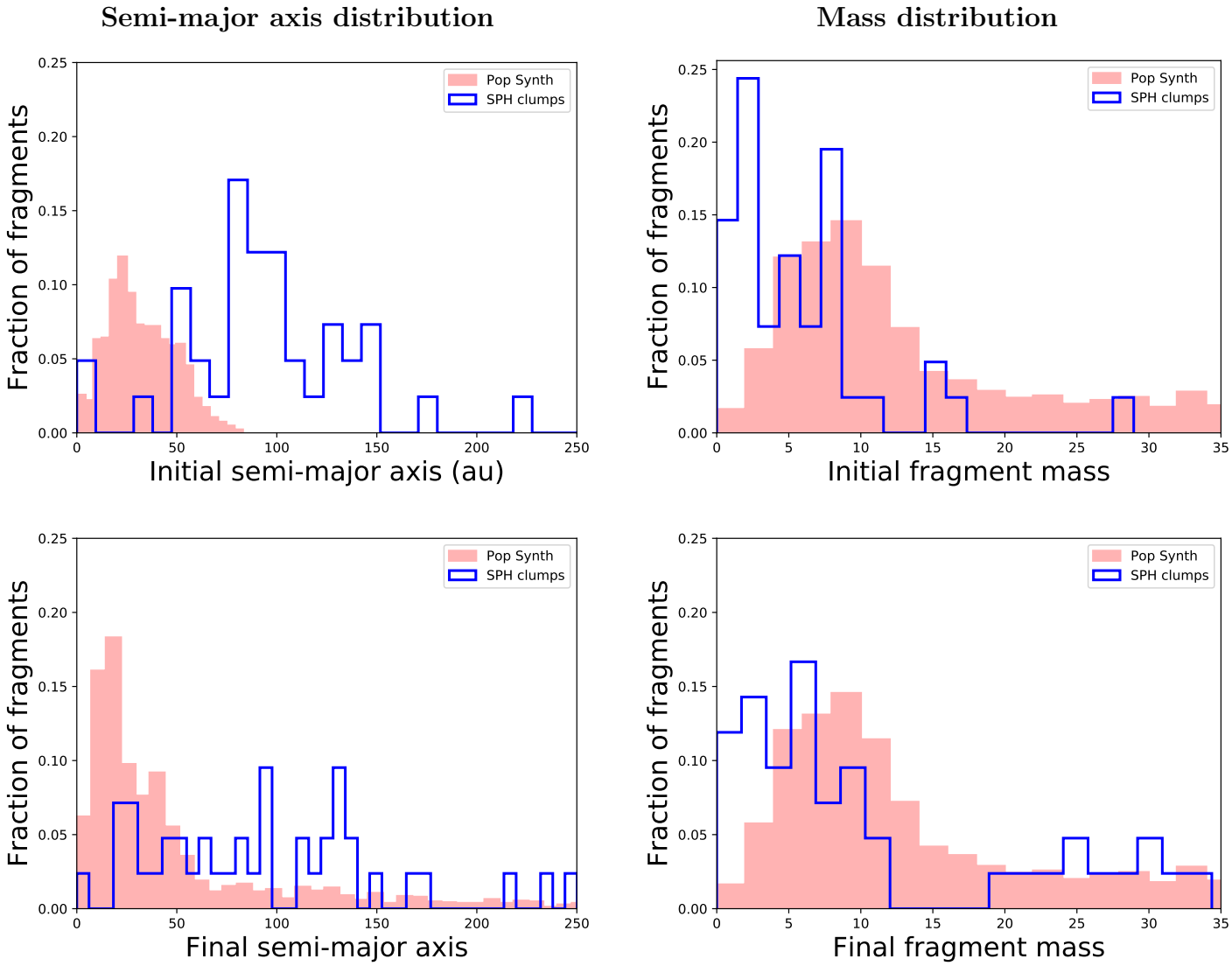


Figure 5.1: Left column shows initial (top) and final (bottom) semi-major axis distribution for our SPH clumps and the population synthesis model of Forgan & Rice (2013b). Right column shows the initial (top) and final (bottom) mass distribution for the same. Population synthesis data are shown in shaded red, while the blue outline shows SPH clumps. Our initial mass and semi-major axis distributions are not strictly accurate, due to limitations of the algorithm requiring a threshold to be met before identification. However, our algorithm is quite robust at later times, and our final semi-major axis distribution shows the importance of clump-clump interactions in the final configuration of a system, with many clumps at large separations due to interactions with each other. Our final masses below 5 M_J are typically underestimated by a factor of 2 what would be identified as mass belonging to the clump "by eye", and accounting for this shows a final mass distribution not unreasonably dissimilar to what we should expect from GI population synthesis models, given the small N statistics we are considering. In all plots, 41 SPH and 1919 population synthesis clumps are plotted.

Having established that the initial mass and semi-major axis distributions for our SPH clumps are subject to some limitations of our detection algorithm, we now compare the final mass and final semi-major axis distributions of our SPH clumps to those in the GI population synthesis model. First, the bottom left hand panel of Figure 5.1 shows a dearth of clumps at $R < 25$ au, when compared to the GIPS model - this is simply due to the last measured value of a before the fragment is destroyed. This figure also shows that the distribution of semi-major axes is very different than is to be expected at this time, given the GIPS model data, and therefore the mechanism that allows these separations to exist at early times (i.e. clump-clump interactions) plays an important role in the ultimate orbital distribution function of the sample.

Second, the bottom right hand panel of Figure 5.1 shows that our final mass distribution is bimodal, with peaks at ~ 5 M_J and ~ 30 M_J. This is somewhat consistent with previous measurements of mass distributions of fragmenting discs by Vorobyov et al. (2013), who find that there are two maxima in their mass distribution, at ~ 5 M_J and ~ 60 M_J. Unlike Vorobyov et al. (2013), we find a gap at ~ 15 M_J, whereas they find a minima at ~ 25 M_J, and our second maxima is at ~ 30 M_J rather than ~ 60 M_J. Given our small N statistics, we would probably expect our distribution to converge on a minima at ~ 15 M_J, rather than the gap that is currently present. Additionally, our second peak is capped at ~ 30 M_J in our simulations since this is typically when the density in a fragment becomes so high that it is computationally unfeasible to continue the simulation. With increased computation time, the mass of our largest clumps would probably increase.

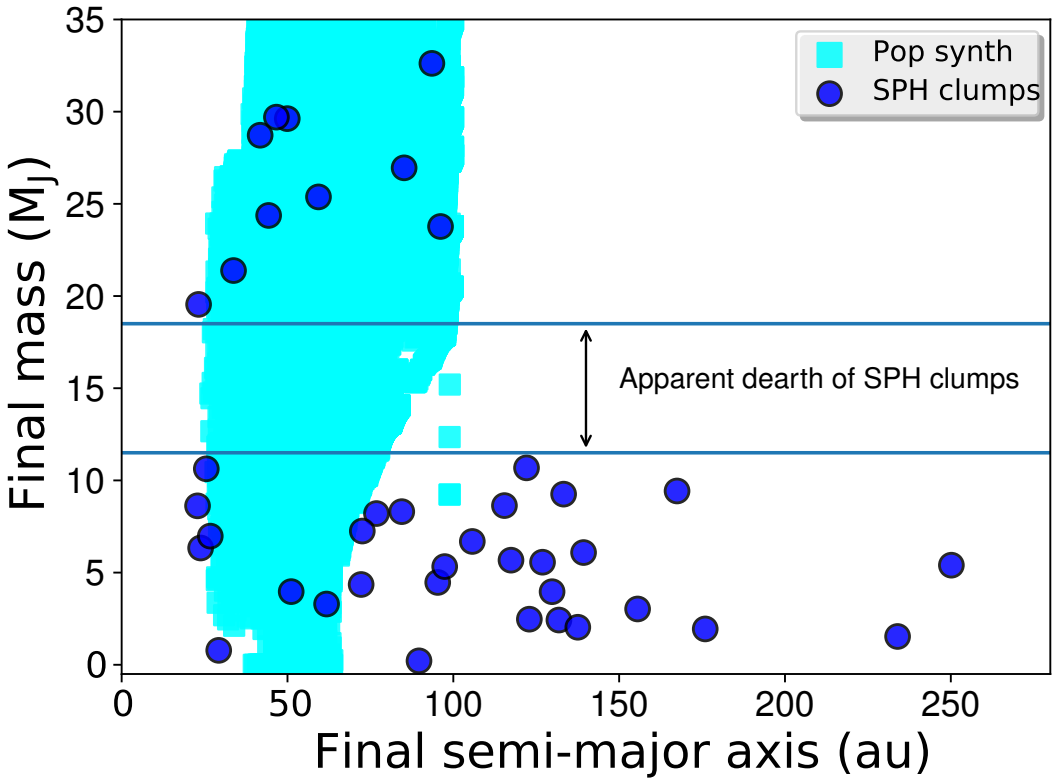


Figure 5.2: Final mass semi-major axis relation for our SPH clumps (dark blue circles) and Forgan & Rice (2013b) fragments (light blue squares). We can see an apparent dearth of intermediate mass SPH clumps at $\sim 10 - 20$ M$_J$. However, since we are dealing with a small sample size we cannot say for sure if this is statistically significant. Despite the small sample size, we can see that large separations for low-mass objects are much more common than suggested by population synthesis models.

Since our algorithm is quite robust at later times, detecting to within a factor of 2 the "by eye" clump mass in low-mass clumps (considered only in the bound region of the clump), we can see, in the final mass distribution shown in the bottom right panel of Figure 5.1, that our GI population synthesis model is significantly underestimating the fraction of planets at < 5 M$_J$, even accounting for under-estimating low mass clumps by a factor of 2.

This can be explained by Figure 5.2, which shows the mass semi-major axis

distribution for the final values of the SPH clumps and the population synthesis fragments. The SPH clumps are the dark circles, and the population synthesis fragments are the light squares. As can be seen, low mass clumps in our simulations are scattered out to large semi-major axis at these early times, and fragment-fragment interactions are likely to play an important role in the ultimate fate of a fragment. If it is scattered out to large a, it is much less likely to be tidally destroyed and far more likely to survive the duration of the simulation. This would suggest that GI population synthesis models need to include fragment-fragment interactions in this early gas phase, since current models suggest that $\sim 40\%$ of initial fragments are tidally destroyed. If a significant fraction of these are scattered out to large radii, their survival rate could potentially be much higher.

5.2.2 Orbital properties

We carry out analysis of the orbital properties of our clumps only using the sample as detected by the density derivative search, as this method is sensitive to most clump masses and semi-major axes. The total semi-major axis evolution of all clumps is shown in the left hand panel of Figure 4.3 on page 167, which we have already discussed, and refer the reader back to. Circles mark surviving clumps (including clumps that subsume another clump), squares mark destroyed clumps, and triangles mark merged clumps. Larger markers correspond to more massive clumps. For destroyed clumps, we take the last measured mass. Roughly half of our most massive clumps migrate radially inwards, which is consistent with migration in locally isothermal discs, as objects exchange angular momentum

with the surrounding gas and move inwards. However, about half of our most massive clumps migrate radially outwards.

This is known to be possible in radiative discs (Kley & Nelson, 2012), but requires either large torques or steep surface density gradients (D'Angelo & Lubow, 2010). For clumps that open a gap, large torques are generated if the gap edge becomes unstable, forcing outward migration through large positive torques (Lin & Papaloizou, 2012). These torques become increasingly positive for more massive clumps (Cloutier & Lin, 2013). Similarly, Stamatellos (2015) found that when a planet opens a gap in a self-gravitating disc, inward migration may be halted or reversed. Although large torques can have many sources, in massive, self-gravitating discs they are likely to be in the form of global spiral arms. We carried out a Fourier analysis on the density structure of our discs, to determine the Fourier amplitude of each m-mode (where m is the number of spiral arms). The amplitude, $A_{\rm m}$, of each mode, m, is calculated by

$$A_{\rm m} = \left| \sum_{i=1}^{N_{\rm region}} \frac{e^{-im\phi_i}}{N_{\rm region}} \right|, \tag{5.6}$$

where $N_{\rm region}$ is the number of particles in the region we are considering (for our case, $R = 20$ to $R = 100$ au), and ϕ_i is the azimuthal angle of the $i^{\rm th}$ particle. Some example amplitudes are shown in Figure 5.3, which shows the first 10 Fourier components of the density structure of 2 discs in their initial state (i.e. when they have just begun to fragment), marked in red, and the same 2 discs in their final state, marked in black. The discs are from simulation 1 and simulation 5, and their final state can be seen in their column density plots, shown

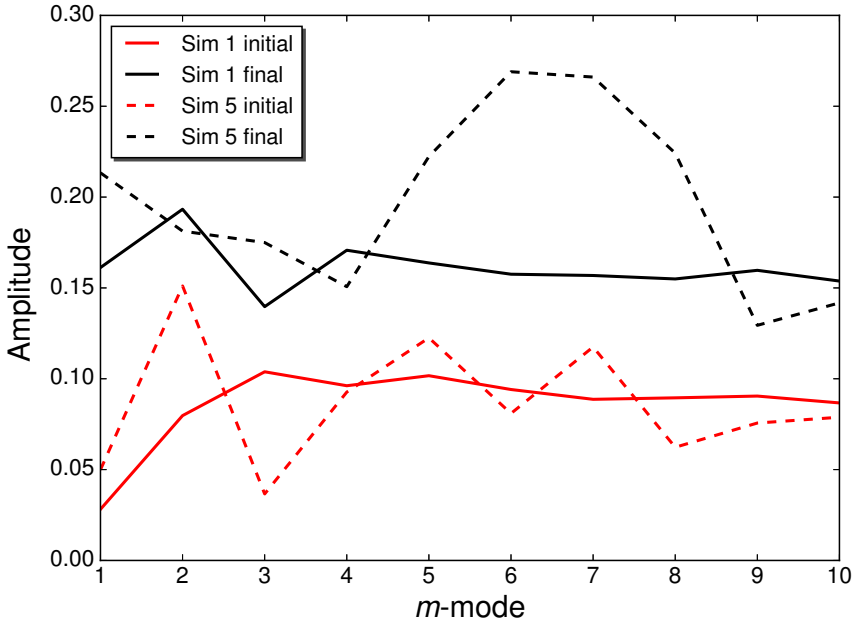


Figure 5.3: Amplitude of the first 10 Fourier components of the density structure of the discs in simulation 1 and simulation 5, calculated between $R = 20$ au and $R = 100$ au, at their initial state (when they have begun to fragment), marked in solid lines, and their final state, marked in dashed lines, at the last timestep.

in Figure 4.1 on page 157. These discs were selected because they ran for the same length of time, and they have contrasting final m-modes states, simulation 1 ultimately peaks in the $m = 2$ mode, and simulation 5 ultimately peaks in the $m = 6$ mode. In this fashion, we determine the dominant final m-mode in each disc, and plot the final semi-major axis of our clumps as a function of this m-mode, in Figure 5.4. We note that our decision to name a dominant mode, based on a relatively low amplitude difference, may be questioned. However, these discs are not in a quasi-steady state, having undergone fragmentation, so persistent spiral modes may be unlikely to form due to tidal disruption from these clumps. Despite the transient nature of the spiral modes, global, low m-mode spiral arms can exert considerable torque, and this is clearly important for the

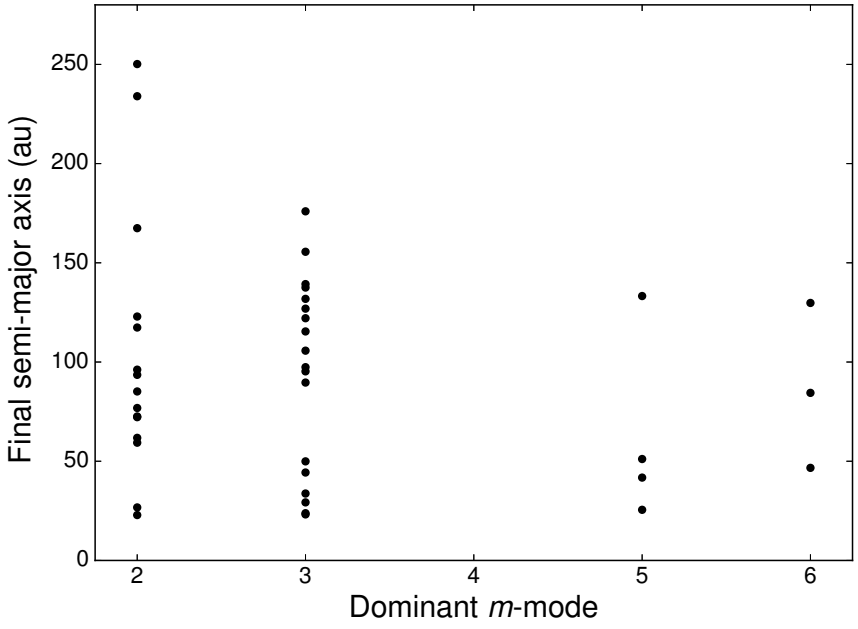


Figure 5.4: Final semi-major axis of all clumps in all simulations as detected by the density derivative search, as a function of the dominant m-mode in the disc. The largest semi-major axis require a 2 armed spiral, which is capable of exerting global torques. There appears to be a rough empirical relationship such that the maximum semi-major axis $a_{\max} \propto 1/m$, although this result is preliminary due to a small number of data points.

final orbital configuration of our clumps, as shown in Figure 5.4, which displays a rough empirical relationship between the maximum semi-major axis of a clump, $a_{\max}$, and m, such that

$$a_{\max} \propto \frac{1}{m}. \tag{5.7}$$

This relationship is one that probably only holds for the largest semi-major axis separation at each spiral mode, since the rest of the points on the plot do not necessarily suggest the same relationship. We discuss our reasoning for the form of this relationship in the conclusion to this Chapter, found in section 5.3.

Of course, this relationship is preliminary, since we only consider 9 discs, all of

the same mass, and it has been shown that more massive discs are dominated by low m spirals (Lodato & Rice, 2004, 2005). Indeed, since it has been shown that in discs without fragmentation we expect the number of spiral modes to be related to the disc-to-star mass ratio, q, such that $m \propto 1/q$ (Dong et al., 2015b), to examine the full parameter space of spiral modes requires a range of q values. Since we consider discs with identical q values, *why* these discs have different dominant m-modes is a valid question. Again, this may be explained by these discs having fragmented into bound clumps. Bound objects in a gaseous disc produce stronger, more persistent torques than spiral density fluctuations alone. If one of our clumps is scattered out of the main body of the disc, say, by interaction with another clump, it may exert a tidal torque on the material in the disc (and by Newton's third law, the material in the disc will also exert a force on the clump). Larger torques are associated with lower m-modes, and if these tidal responses from the discs to the clumps is responsible for the low m-mode domination in the disc, then we expect to see steeper surface density profiles in discs with low m-modes, since more mass will have been redistributed inwards as a result of this torque. This is difficult to unequivocally demonstrate in our set of simulations, since each simulation was run for a different length of time, and the amount of mass redistributed increases with time.

However, Figure 5.5 shows the mass enclosed as a function of radius for the two discs plotted in Figure 5.3, that ran for the same length of time. Although by no means conclusive, the slight increase in mass for a given radius between 40 au and 160 au for the $m = 2$ disc in simulation 1 is consistent with tidal forces being responsible for the low m-mode becoming more dominant. To properly

establish the nature of the preliminary relationship detailed in equation 5.7 therefore requires a range of disc masses and fragmentation scenarios, and we leave this to future work.

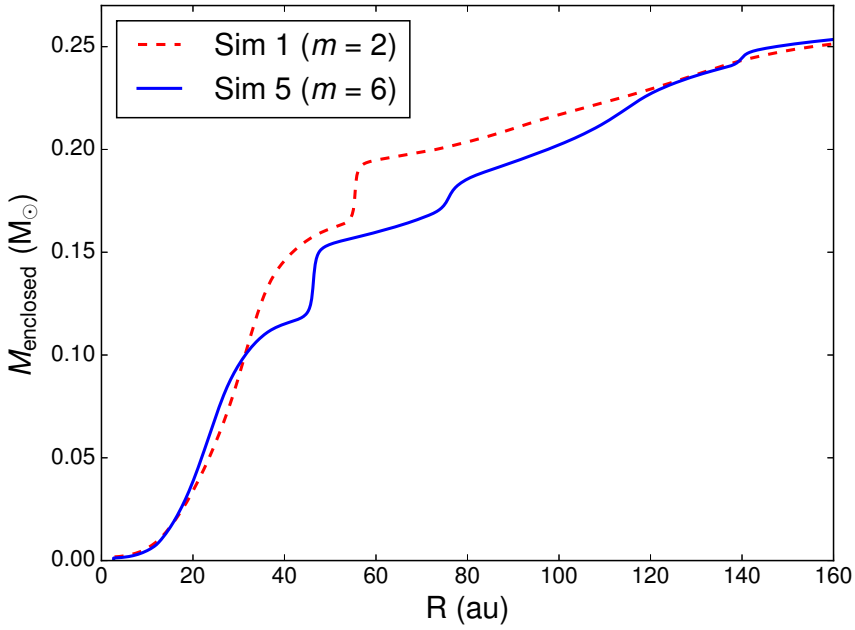


Figure 5.5: Disc mass enclosed as a function of radius for the final timesteps of simulation 1 (red dashed line) and simulation 5 (blue solid line). Since more mass is enclosed at shorter radii for simulation 1, more mass has been transported inwards in the disc, giving it a slightly steeper density profile.

Figure 5.6 shows the relationship between eccentricity, e, and semi-major axis for our SPH clumps. Larger markers indicate more massive clumps. For the most part, the more massive clumps are located on close in ($a \sim 50$ au), low eccentricity ($e \sim 0.1$) orbits, while lower mass clumps are at larger separations and with higher eccentricity. Since disc fragmentation forms objects on low eccentricity orbits ($e < 0.1$), we can see the importance of clump-clump interactions in determining the final orbital properties of a clump. Very large eccentricities ($e \sim 0.7$) at large a indicate that a clump is close to ejection, as excitations beyond unity ensure a

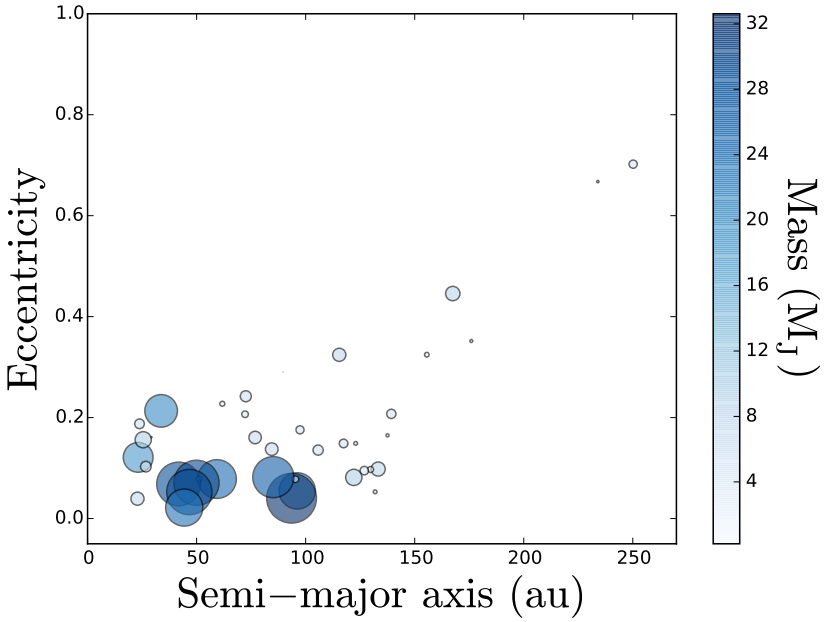


Figure 5.6: Final semi-major axis and eccentricity relation for our SPH clumps. Larger markers represent more massive clumps. The population synthesis model does not contain eccentricity data. We can see, for the most part, our clumps have a roughly linear relationship between eccentricity and semi-major axis, although there is a large amount of scatter.

clump is ejected from the disc.

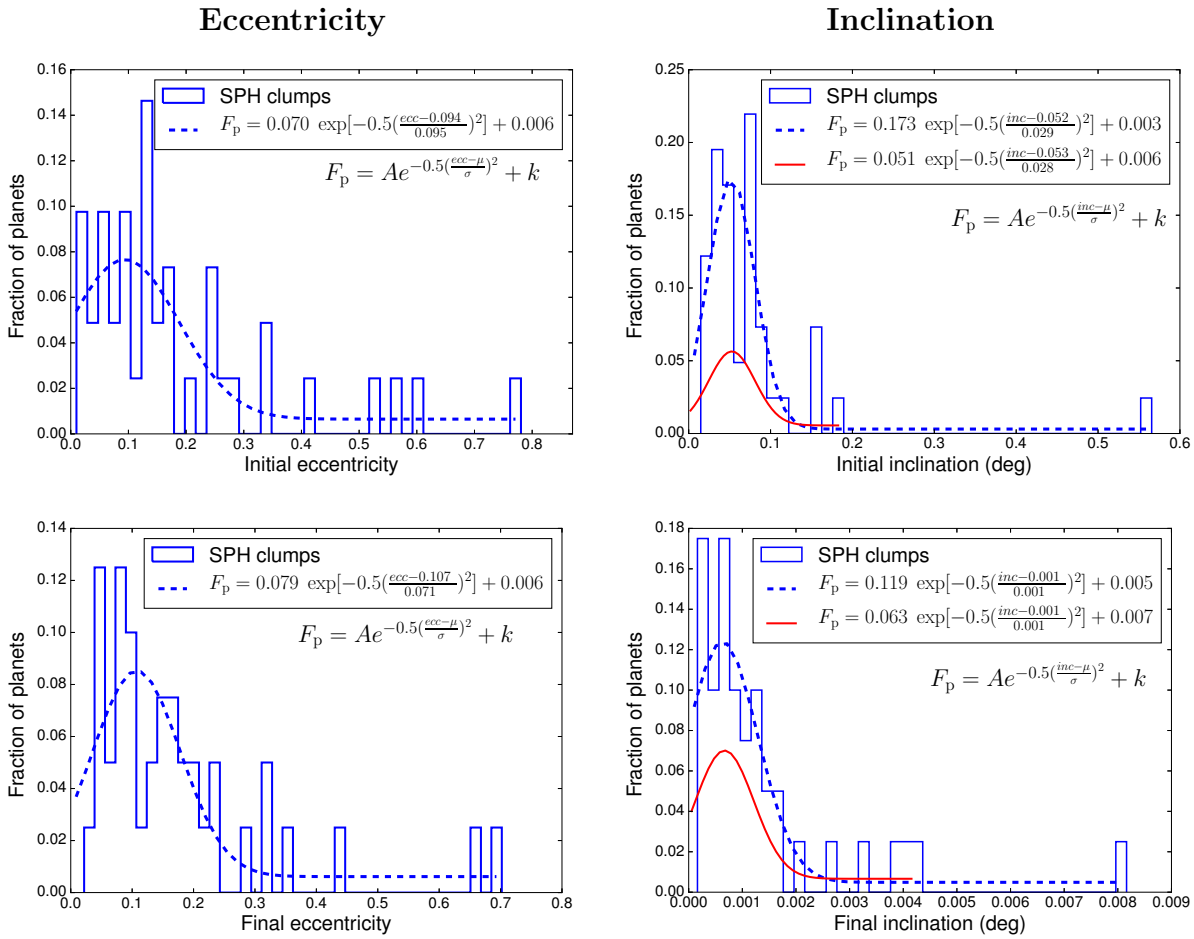


Figure 5.7: Left column shows initial (top) and final (bottom) eccentricity distribution for the 41 SPH clumps formed in the 9 SPH simulations. Right column shows the initial (top) and final (bottom) inclination distribution for the same. Inclination is calculated relative to the orbital plane of the central star. Since the population synthesis models of Forgan & Rice (2013b) do not contain eccentricity or inclination information, we do not plot them here. Using least-squares regression, we have fitted our distributions with a Gaussian of the form $F_p = A\exp[-0.5(\frac{x-\mu}{\sigma})^2] + k$, where F_p is the fraction of planets, A is a fitted constant, x is eccentricity or inclination, μ is the mean of the distribution, σ is the standard deviation and k is the offset constant. The fitted values are given in Table 5.1 and in each plot legend. We have included these fits since we consider they may be useful in developing future GI population synthesis models, but caution that our sample size is small. Our inclination histograms have been fitted with two distributions. The dotted blue line includes all data points, and the solid red line does not include the most inclined point in each distribution, since with a small sample size, and an apparent gap between the rest of the clumps, it is unclear whether or not this is an outlier. There is little change between our initial and final eccentricity distribution, both peaking at $e \sim 0.1$, and a slightly smaller standard deviation for the final configuration. The inclinations of our clumps are decreased by a factor of ~ 100 between their initial and final states, showing that clump orbital inclination is rapidly adjusted after formation, until it orbits almost entirely in the plane of motion of the central star.

The top two panels of Figure 5.7 show the initial eccentricity distribution (left) and initial inclination distribution (right) of our SPH clumps. The bottom two panels of Figure 5.7 show the final eccentricity distribution (left) and final inclination distribution of the same. Inclination is calculated relative to the orbital plane of the central star, such that

$$i = \arccos\left(\frac{L_{\rm z}}{|\vec{L}|}\right), \tag{5.8}$$

where i is the orbital inclination of the clump, $\vec{L}$ is the orbital angular momentum vector of the clump (calculated relative to the centre of mass and centre of velocity of the central star), and $L_{\rm z}$ is the z component of $\vec{L}$. Using least-squares regression, each plot has been fitted with a Gaussian of the form

$$F_{\rm p} = Ae^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} + k, \tag{5.9}$$

where $F_{\rm p}$ is the fraction of planets, A is a fitted constant, x is either eccentricity or inclination, μ is the mean of the fitted distribution, σ is the standard deviation of the fitted distribution and k is the fitted offset constant. The fitted values are given in Table 5.1, and in the legend of each plot. The initial and final inclinations have been fitted with two distributions, the dotted blue line includes all points, and the solid red line does not include the most inclined point in each distribution, since there is a large gap between that point and the rest of the clumps, and a small sample size, it is unclear if this point is actually an outlier. We have provided these fits since current GI population synthesis models do not include orbital eccentricity or inclination, and, despite our small sample

size, this information may be useful in the future development of these models. Aside from a small decrease in standard deviation, there is little change between our initial and final eccentricity distribution; both peak at $e \sim 0.1$ and share an offset constant of $k = 0.006$. However, the orbital inclination of our clumps is reduced by a factor of ~ 100 between the initial and final states, showing that clump orbital inclination, in our SPH simulations, is rapidly reduced after formation. Considering that many of our clumps undergo dynamical interactions that cause scattering and eccentricity pumping on short timescales, this high degree of coplanarity may be surprising, especially when considering that most exoplanets have mutual inclinations of a few degrees (Fang & Margot, 2013; Figueira et al., 2012). However, it is consistent with our current understanding of highly inclined planet orbits relying on dynamical perturbations such as the Lidov-Kozai mechanism (Naoz et al., 2013). Our results may indicate that developing inclined orbits is difficult while a gas disc is present, even if substantial dynamical interactions between clumps take place in this time.

5.2.3 Spin properties

We analysed all of the fragments in our simulations, and found that several of them survive to the end of the simulation whilst undergoing retrograde rotation. This is shown in Figure 5.8, which shows the relative alignments between the orbital angular momentum vector and the rotational angular momentum vector of the clumps. Both the top and bottom panel is split into two parts, showing significant misalignment at the top, marked in red crosses, and good alignment at the bottom. The left panel shows the alignment as a function of mass, and

Plot	μ	σ	A	k
Initial ecc.	0.094	0.095	0.070	0.006
Final ecc.	0.107	0.071	0.079	0.006
Initial inc. (no outlier)	0.052	0.029	0.173	0.003
Initial inc. (with outlier)	0.053	0.028	0.051	0.006
Final inc. (no outlier)	0.001	0.001	0.119	0.005
Final inc. (with outlier)	0.001	0.001	0.063	0.007

Table 5.1: Parameter values of the Gaussian fits applied to the histograms in Figure 5.7. From left to right, the columns are plot, mean, standard deviation, amplitude (without offset) and offset constant.

the right panel shows the alignment as a function of semi-major axis.

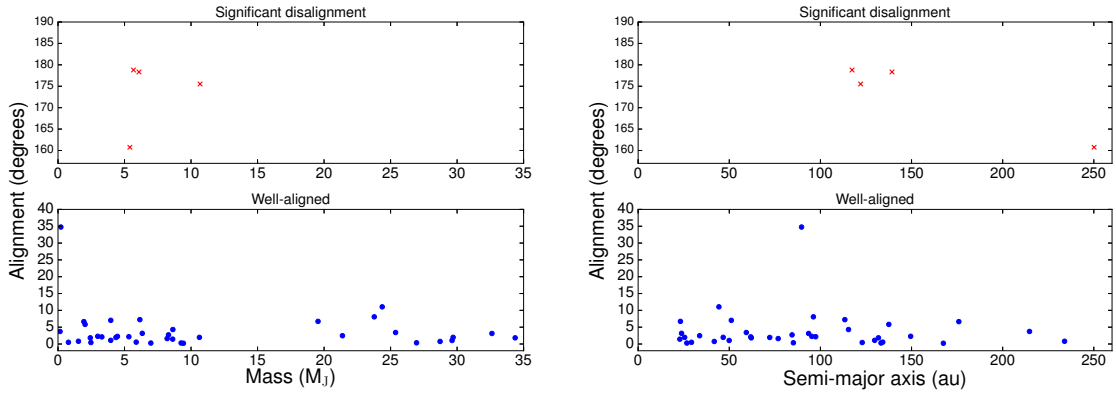


Figure 5.8: All plots show alignment, on the y-axis, defined as the angle between the orbital angular momentum vector and the rotational angular momentum vector. Left two panels show alignment as a function of mass, right two panels show alignment as a function of semi-major axis. The plots are split, for clarity, into two different regions, between 0^o and 40^o is on the bottom, and between 155^o and 190^o is on the top. Well aligned clumps are marked by blue circles, clumps with a significant degree of misalignment are marked with red crosses. Both plots show that four low-mass, high-separation clumps are retro-rotators, given that all of them or orbiting in prograde motion.

This prompts the question - how did they get to be retro-rotating? Did they form

like this, or were they perturbed in some way? Having checked all of the clumps with significant misalignment, we can see that all of them were perturbed by a close encounter with another fragment, which typically flung them rapidly further out into the disc. We show the most extreme example in Figure 5.9, which shows the retrorotating clump 4 in simulation 4. In the leftmost panel, we see in the bottom left corner two clumps undergoing a close encounter. One of them then decreases its semi-major axis, whilst the other one is flung further out into the disc, to become retrorotating.

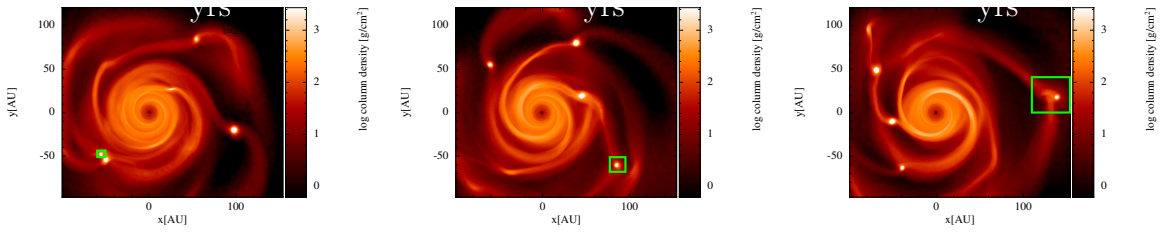


Figure 5.9: Column density plots of simulation 4, increasing in time from left to right, showing clump 4 (marked by green square), an initially prograde-rotating clump, undergoing an encounter with another clump to become a retro-rotating clump.

Figure 5.10 shows the specific angular momentum profiles of a retro-rotating clump (left) and a prograde rotating clump (right). The blue line, at $T = 2788$ years, is as soon is the clump is detected by our algorithm. We can see that the majority of the clump is in prograde rotation, with only the outer ~ 1 au in retrograde rotation. However, as time progresses and the clump continues to accrete material that is retro-rotating due to the encounter, this ultimately changes the rotation of the whole clump. For comparison, the right panel contains

a prograde rotator of similar mass from simulation 5.

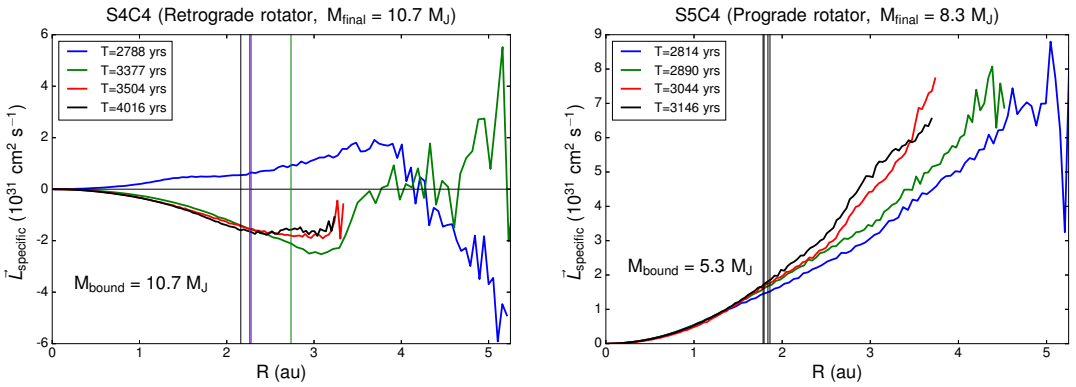


Figure 5.10: Left: radial profile of specific angular momentum for an example retro-rotating clump (clump 4, simulation 4), at four different times. Vertical lines indicate the last bound point of the clump at times indicate in the legend. The final bound mass is marked on the plot, and in this case, all of the mass that is identified as belonging to the clump is ultimately bound to the clump. Positive values for $\vec{L}_{\rm specific}$ indicate that the rotational angular momentum vector of the material and the orbital angular momentum vector of the whole clump are aligned (or at least inclined at less than 90°), and negative values indicate that the two vectors are anti-aligned. The blue line is when the clump is first detected, and we can see that the majority of the clump is in prograde rotation, and the outer ~ 1 au is in retrograde rotation. The material came to be retro-rotating due to a close encounter with another clump, which is shown in Figure 5.9. As the clump continues to accrete material, we can see that angular momentum is exchanged between the inner material and the outer material (green line, $T = 3377$ years). As the clump contracts, this positive angular momentum material is no longer considered part of the clump. Right: for comparison, the specific angular momentum profile of a clump of comparable mass undergoing prograde rotation (clump 4, simulation 5). The vertical line indicates the last bound radius in the clump for the four clumps, only one line is plotted here as the four are so close together. The final bound mass of the clump is stated on the plot.

Figure 5.11 shows the rotation velocity profiles of two clumps. The left-hand panel shows the ultimately retro-rotating clump 4 in simulation 4, and the right-hand

panel, for comparison, is a clump of comparable mass that is always undergoing prograde rotation. Negative velocity is determined by the orbital angular momentum vector of the whole clump and the rotational angular momentum vector of the material being anti-aligned. Both panels show the clumps at four different times, and the dashed lines indicate the breakup velocity profile of each clump. Both clumps are rotating under their breakup velocity for radii below 1 au, and their velocity profiles are consistent with solid body rotation (i.e. $v \propto R$) at these radii. Much further out, the clumps become nebulous, but we have included information out as far as possible to show the interesting angular momentum exchange between material at $T = 2788$ years and $T = 3377$ years for the retro-rotator.

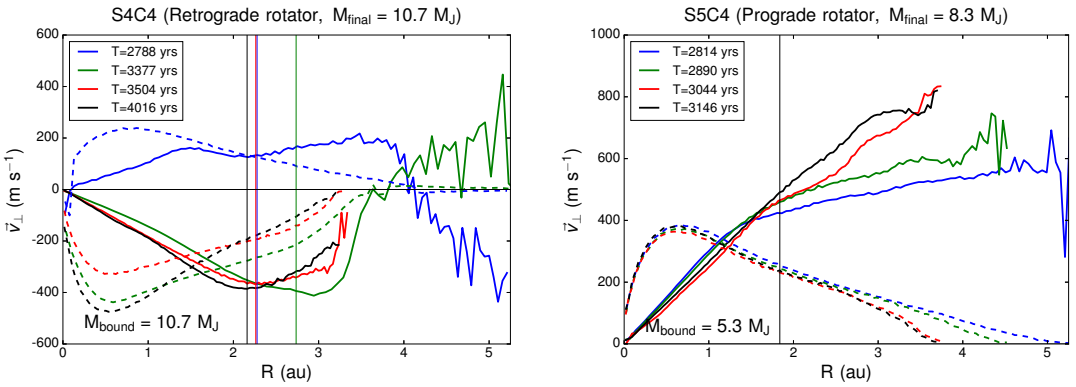


Figure 5.11: Left: rotation velocity curve for an example retro-rotating clump (clump 4, simulation 4), at four different times. Vertical lines mark the last bound radius in the clump at times given in the legend. In this case, the total bound final mass is equal to the total mass ultimately identified for the clump. Dashed lines correspond to breakup velocities at these times. Negative velocities indicate that the rotational angular momentum vector of the material and the orbital angular momentum vector of the whole clump are anti-aligned (or at least inclined at more than 90°) relative to each other. We can see that beyond $\sim$ 1.5 au, the material is rotating at faster than its breakup velocity, which would perhaps suggest that material is spreading outwards from the clump, in a disc-like, or toroidal, manner. Right: for comparison, a clump of comparable mass undergoing prograde rotation (clump 4, simulation 5). Again, beyond $\sim$ 1 au, rotation velocity exceeds breakup velocity, and so we may expect to see a considerable spread of material around such an object, morphologically similar to a toroid.

For the prograde rotator, the ultimate configuration is a good approximation to a solid body rotation curve out to extended radii ($\sim$ 3 au). Interestingly, for the retro-rotator, the velocity profile at the outer part of the fragment ($\sim$ 2.2 au to $\sim$ 4 au) is consistent with Keplerian rotation (i.e. $v \propto 1/\sqrt{R}$). This suggests the presence of a disc, or a disc-like structure, around the clump. Unfortunately, it is not (at the time of writing) currently possible to self-consistently re-resolve such regions in SPH simulations, so we are unable to investigate this further.

5.2.4 Density and temperature properties

Maximum density and temperature for all clumps in our simulation is shown in Figure 5.12, open circles show prograde rotating clumps and dark triangles show retrograde rotating clumps. Only one clump out of the 41 present in our simulations reaches a temperature above the 2000 K threshold necessary to dissociate molecular hydrogen, with a temperature of 2081 K. This implies that all clumps, except for this one, may be considered as the first hydrostatic cores.. Additionally, a further 9 clumps have internal maximum temperatures above $\sim$ 1000 K, which means they would begin to evaporate dust.

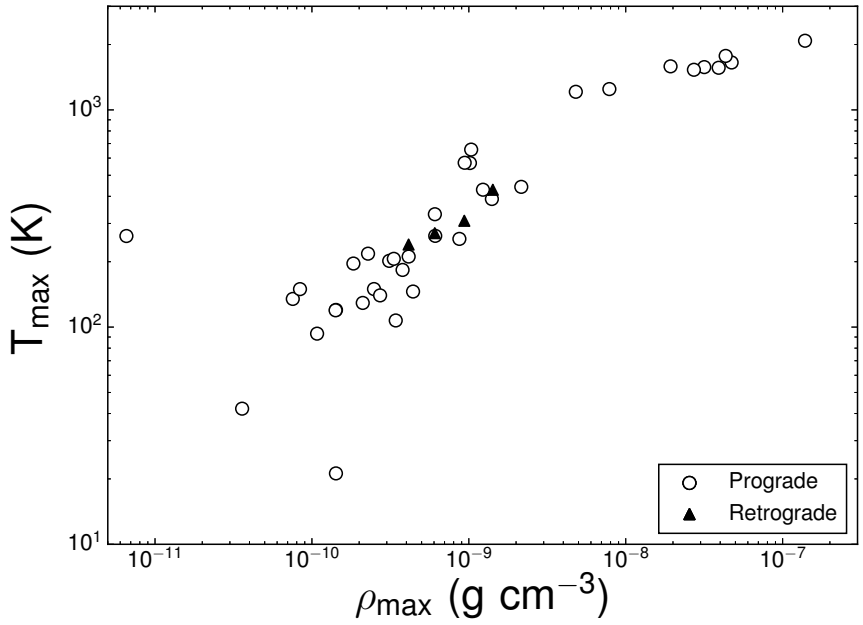


Figure 5.12: Final maximum temperature and density of clumps in all simulations. Open circles are prograde rotating clumps, closed triangles are retrograde rotating clumps.

Both of these results are in good agreement with previous studies of fragments, such as Vorobyov et al. (2013), who found similar results. Our measurements

of clump maximum temperature are limited by our resolution, since running simulations at higher resolution would allow higher densities to be reached before becoming computationally infeasible to continue.

As discussed in the introduction to this chapter, the advantage of using a purely hydrodynamical simulation with no sink particles is that we can examine fragment internal structure during the simulation. With this in mind, we show the radial temperature and density profiles of 7 fragments, one of which ultimately becomes the hottest fragment in our simulations, shown in Figure 5.13, and the remaining 6 are three retro-rotators and three prograde rotators, of comparable mass and with similar simulation histories, shown in Figure 5.14. In both figures, each image is split into two panels, radial temperature profile on the top and radial density profile on the bottom.

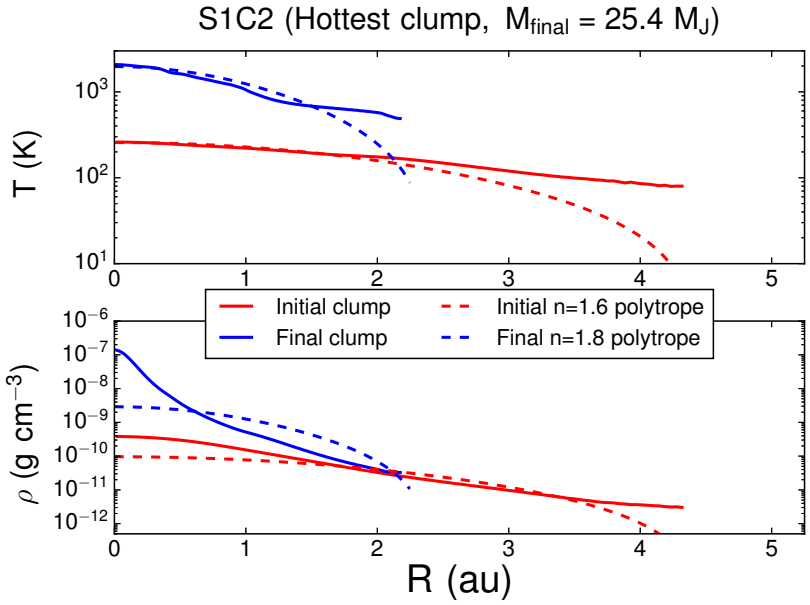


Figure 5.13: Radial temperature and density profile of the hottest clump we identified in our simulations, clump 2 in simulation 1. Red solid lines show initial temperature and density, blue solid lines show final temperature and density. For comparison, an $n = 1.6$ polytrope is plotted for the initial clump, and an $n = 1.8$ polytrope is plotted for the final clump. In both cases, the polytrope is a good fit to the temperature profile of the clump, however it is a poor fit to the density, particularly in the final state, where it underestimates central density by almost two orders of magnitude.

The initial clump profile is shown in solid red, and the final clump profile is shown in solid blue. For comparison, a polytrope (dashed line) is also plotted in each panel. Figure 5.13 shows the radial density and temperature profile of the hottest clump identified out of our 9 simulations. We can see that for the initial state, the radial temperature profile is reasonably well described by a polytrope, of index $n = 1.6$, out to a radius of ~ 2 au. However, the final temperature and initial and final density all deviate significantly from the polytropic value.

The rest of the clumps, shown in Figure 5.14, show reasonable temperature agreement with polytropes out to radii of ~ 2 au, however none of them fit

particularly well to density. The left-hand column shows prograde rotators, and the right-hand column shows retrograde rotators. In all cases in Figure 5.14, we have plotted a polytrope of index $n = 1.5$, which is appropriate for a fully convective star such as a brown dwarf. Since in a polytrope, pressure P and density ρ are related by

$$P = K\rho^{\frac{n+1}{n}}, \tag{5.10}$$

where K is a constant and n is the polytropic index, it implicitly assumes that pressure is a power law function of density which is constant throughout the star. For our clumps, it appears to be the case that a polytropic approximation may be too simplistic when estimating the internal structure of the clump. This may have some implications again for current GI population synthesis models, since orbital parameters are sensitive to the radial distribution of mass in the forming fragments.

When we compare the final states of the retro-rotating clumps with their prograde rotating counterparts, there is not much that would mark them as retro-rotating, final density and final temperature profiles for both directions of rotation are similar, and are consistent with other simulations of fragmenting protostellar discs (see, e.g. Vorobyov et al. 2013).

There is one clump that is an exception when compared to the rest of the clumps, and is shown in the top right hand panel of Figure 5.14. This is S4C4. In its final state, the surface temperature of the clump is a factor of ~ 3 higher than we see in the rest of the clumps, and its central temperature is ~ 425 K, a factor of ~ 2 larger when compared to the rest of the clumps in Figure 5.14. This high

temperature could possibly be explained by an encounter with another clump. Figure 5.9 shows this interaction, (clump 4 is marked by a green square). It is scattered by the more massive clump into the outer part of the disc, becoming shocked as it passes through a spiral arm. The encounter with the other clump causes the direction of rotation of clump 4 to change, but the large increase in temperature could be due to this motion through a region of increased density, entering perpendicular to the spiral arm where the density gradient is at its

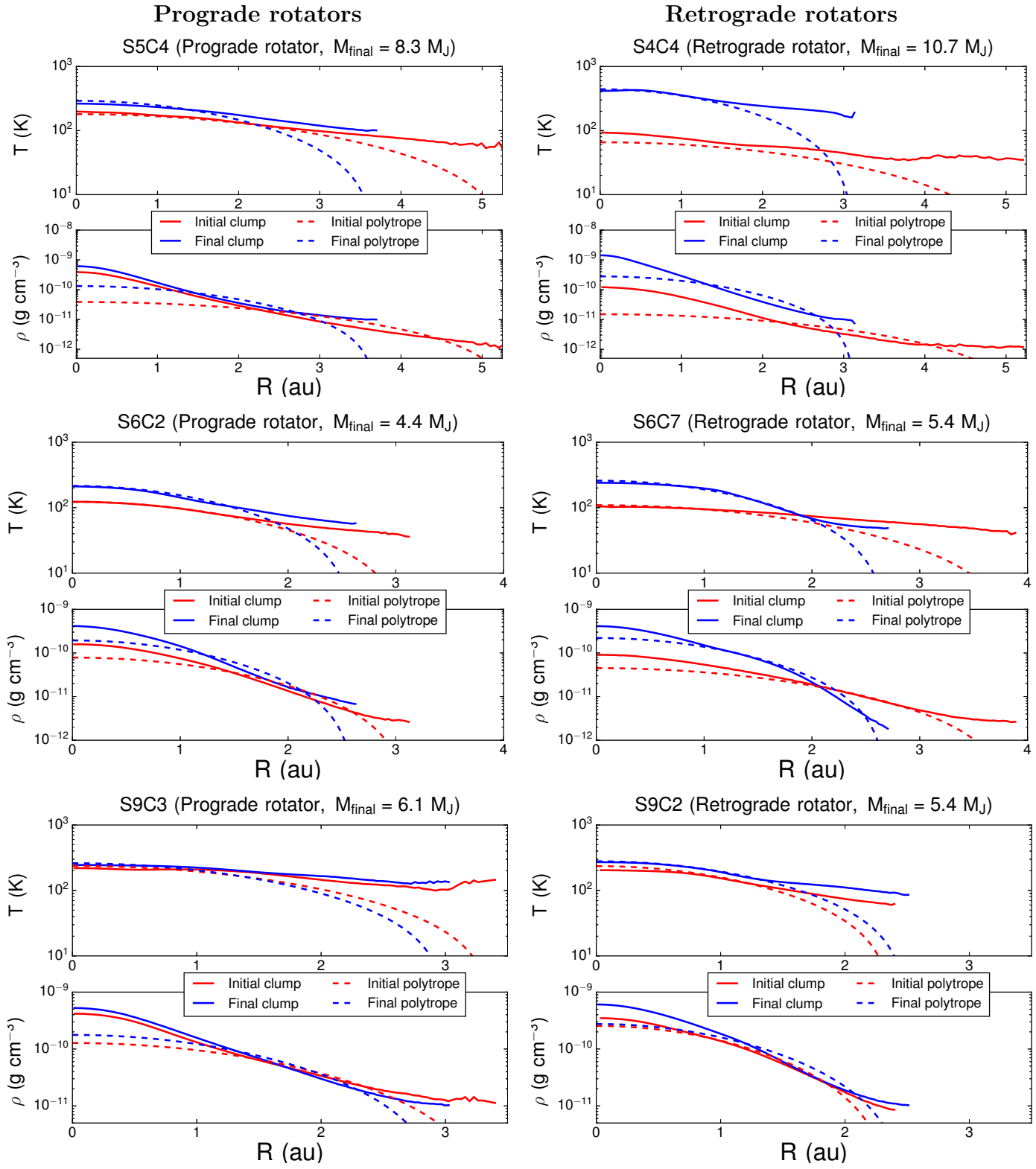


Figure 5.14: Radial density and temperature profiles for the initial and final state of 6 clumps. Red solid lines show initial clump profile, blue solid lines show final clump profile. For comparison, all clumps have initial and final $n = 1.5$ polytropic profiles plotted in initial and final states. The left column contains 3 clumps undergoing prograde rotation, and the right column shows 3 clumps undergoing retrograde rotation. In each row, the clumps are of comparable mass, and are intended to be compared, although bearing in mind that the clumps have had different evolutionary histories and their final masses are not identical. For most of the clumps, there is not much in their profiles that would mark them as a retro-rotating clump. The final inner and surface temperatures are similar, as is the shape of the density profile. Clump 4 in simulation 4 (top right panel) is somewhat the exception, with a large increase in both final inner density, inner temperature and surface temperature. Given the violent encounter it endured early in its history (shown in Figure 5.9), this may be unsurprising, but it is interesting to note that high surface temperatures may indicate a violent encounter in the past. In all cases, a polytrope of index $n = 1.5$ is a reasonable fit to the temperature profile, but consistently underestimates the inner density by around an order of magnitude.

largest. We suggest, therefore, that clump-clump interactions may provide a mechanism for dramatic increase in temperature of forming clumps, either directly through the interaction, or their subsequent scattering through dense regions in the disc. This mechanism could present a problem for terrestrial planet formation through tidal downsizing. For a terrestrial planet to form in the tidal downsizing scenario, a clump must not accrete too much mass, and then become tidally disrupted after migrating too close to the host star (Boley et al., 2010; Nayakshin, 2010a). It then leaves behind a solid core, *if* dust grain sedimentation was sufficiently rapid to form a core. Since we have tentatively shown that clump-clump interactions are common (given our small sample size), then it may be possible that clump temperatures are frequently too high, at too young an age, for dust sedimentation to have taken place inside clumps. For a solid core to form, clumps need to exist at a temperature below the dust sublimation temperature ($\sim$ 1200 K) for the duration of the sedimentation process.

5.2.5 Tidal disruption and mergers

Of the 41 clumps detected by the DDS method in our simulations, 7 were tidally destroyed by the central star, and 4 clumps underwent mergers. Despite our small sample size, these results suggest that both tidal destruction and mergers are common amongst protostellar disc fragments. An example merger is shown in Figure 5.15, which shows simulation 3, clumps 2 and 4, merging together.

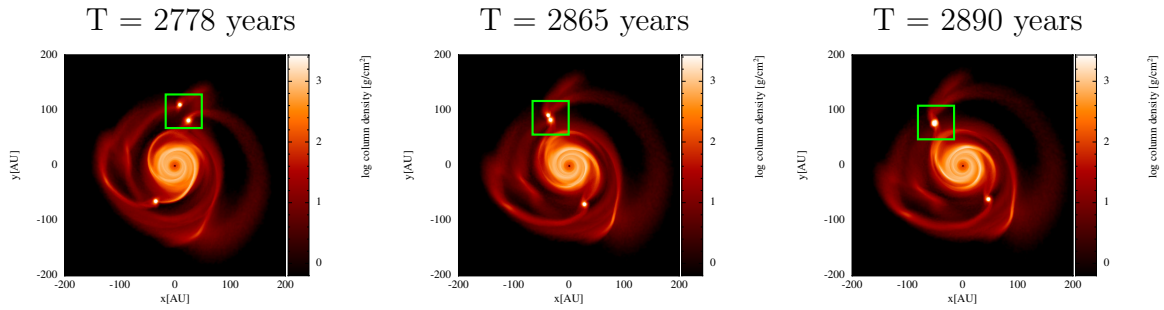


Figure 5.15: Column density plots of simulation 3, where clumps 2 and 4 (highlighted in green) undergo a merger.

An example tidal disruption is shown in Figure 5.16, which shows simulation 7, clump 4 undergoing tidal disruption. Such tidal disruptions could potentially be an explanation for outburst type behaviour in young protostars, due to the rapid increase in accretion rate onto the central star (see, e.g.,Boley et al. 2010; Nayakshin & Lodato 2012; Vorobyov & Basu 2005). We leave observational signatures of our tidal disruptions to future work.

5.3 Discussion and Conclusion

In this chapter, we have built on the work of Chapter 4, making use of our novel algorithm, the density derivative search, to detect and trace fragments for the duration of the simulation. We ran 9 SPH simulations of a 0.25 $M_\odot$ disc around a 1 $M_\odot$ star, each with a surface density profile of $\Sigma \propto r^{-1}$, an inner radius of 10au and an outer radius of 100 au. Each simulation was run for as long as computationally feasible without converting dense regions to sink particles, since we wished to calculate orbital properties for our fragments, which are sensitive

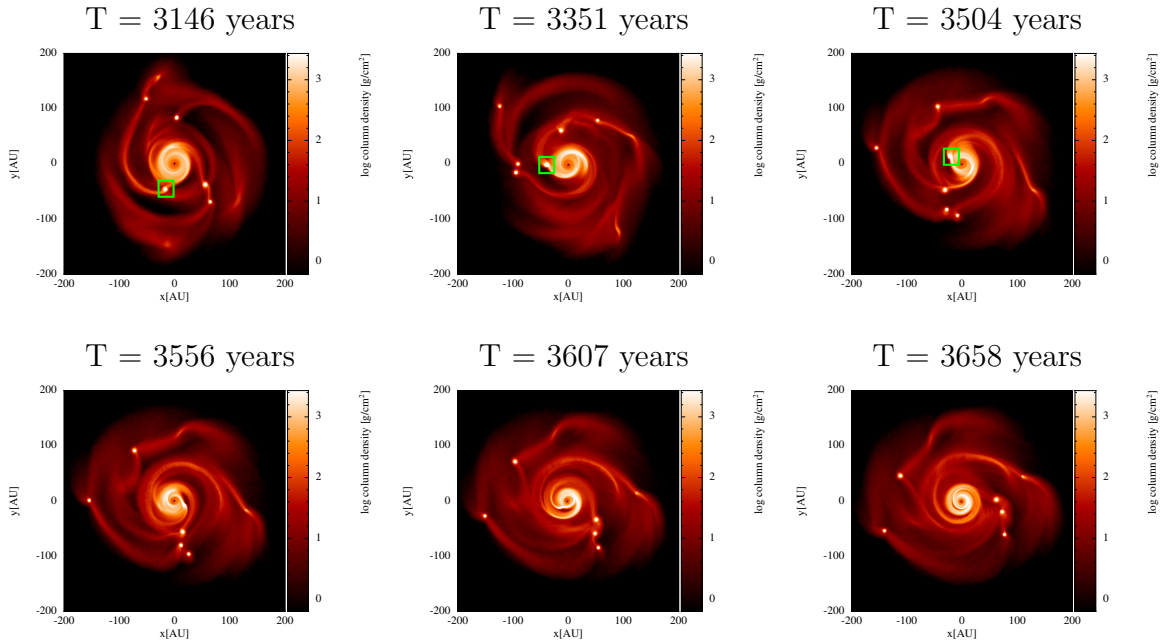


Figure 5.16: Column density plots of simulation 7, increasing in time, showing the tidal disruption of clump 4, marked in green in the top three images before it begins to be tidally destroyed.

to their radial mass distribution.

We compare our sample of fragments to the population synthesis model of Forgan & Rice (2013b), and find that our algorithm has some limitations at early times (i.e., during the initial period of fragment formation), which means that some radial migration has already happened before the algorithm detects the fragment. Despite this, it is fairly robust at late times (i.e., a few orbital periods after formation), and so our final mass and final semi-major axis functions, and the mass semi-major axis relationship are representative of the final configurations of our systems.

We examine the internal temperature and density structure of our fragments, and compare them with appropriately indexed polytropes. We find that the central density of our fragments are typically an order of magnitude denser than their polytropic equivalent, and since orbital parameters of a body are sensitive to its internal mass distribution, we recommend caution when using polytropes in GI population synthesis models when calculating orbital parameters. Additionally, we find that our fragments may be shocked via spiral arm passage through interactions with each other, rapidly increasing the internal energy of the fragment. This could have implications for terrestrial planet formation through the downsizing hypothesis, since solid core formation requires rapid dust sedimentation. If the interior of these fragments are hot enough to sublimate dust at very early stages in their lifetime, it may not be possible for them to form solid cores.

We find that fragment-fragment interactions play a substantial role in the

ultimate fate of our systems. Low-mass fragments can be scattered out to large radii, (and therefore remain low mass, since there is less material to accrete), and are therefore unlikely to be tidally destroyed by the central star. Since current GI population synthesis models suggest that $\sim 40\%$ of initial fragments are ultimately tidally destroyed (Forgan & Rice, 2013b), if a significant number of these fragments are actually scattered out to large separations by interactions with other fragments, this figure could potentially be much lower. We therefore recommend that fragment-fragment interactions in the gas phase of the disc be included in any new GI population synthesis models. In addition to encounters scattering fragments out to large radii, we also find a tentative relationship between the dominant azimuthal wavenumber in the disc, and the maximum semi-major axis of a clump in that disc, such that $a_{\max} \propto {}^{1}/_{m}$. Although this relationship is preliminary, and requires more simulations in a range of disc-to-star mass ratios to confirm it (since $m \sim {}^{1}/_{q}$ (Dong et al., 2015b), where q is disc-to-star mass ratio), it is unsurprising that such a relationship may exist, and we guess that the relationship is one of inverse proportionality, rather than one of negative proportionality, due to the relationship between gravitational torque $\Gamma_{\rm G}$ and azimuthal wavenumber as follows (for a full derivation and comprehensive explanation, we refer the reader to Binney & Tremaine 2008). The gravitational torque $\Gamma_{\rm G}$ exerted on material outside a radius R_0 in a disk is given by

$$\Gamma_{\rm G} = \mathrm{sgn}(k)\frac{\pi^2 m R_0 \mathrm{G}\Sigma_1^2}{k^2}, \tag{5.11}$$

where Σ_1 is a gentle function of radius, k is the radial wavenumber defined as

$$k \equiv \frac{\partial(mf(R))}{\partial R}, \tag{5.12}$$

$mf(R)$ is the radial shape function of the spiral (for a simple example, see Hall et al. 2016), and $\mathrm{sgn}(k) = +1$ for trailing spirals (i.e, positive torque exerted outwards). At a given value of R_0, and the same function for Σ_1, we can see that

$$\Gamma_{\mathrm{G}} \propto \frac{1}{m}. \tag{5.13}$$

To establish our relationship $a_{\mathrm{max}} \propto {}^{1}/_{m}$, we have assumed that the amount of torque is directly proportional to the change in radial distance. Whilst this is almost certainly an over-simplification of matters, we can see that from

$$\begin{aligned} \Gamma_{\mathrm{G}} &= \frac{d|\vec{L}|}{dt} \\ &= \frac{d}{dt}\Big(m|\vec{v} \times \vec{r}|\Big) \\ &= \frac{d}{dt}\Big(m|\vec{v}||\vec{r}|\sin[\theta]\Big), \end{aligned} \tag{5.14}$$

if we assume mass, velocity, and the angle between $\vec{v}$ and $\vec{r}$ stay fairly constant, then $\Gamma_{\mathrm{G}} \propto {}^{dr}/_{dt}$, i.e. the distance we wish to move our fragment. Assuming that all fragments form at roughly the same r in a given disc, (i.e. where that particular disc becomes susceptible to fragmentation), then we recover $a_{\mathrm{max}} \propto {}^{1}/_{m}$.

That $\Gamma_{\mathrm{G}} \propto {}^{1}/_{m}$ could potentially impact the fragmentation radius in a given disc is unsurprising. For low m, large torques will be exerted on the disc, allowing mass to move inwards towards the central star. If the surface density profile of the

disc changes significantly, then the radius at which the disc becomes susceptible to fragmentation will also change. With this in mind, we recommend that future GI population synthesis models have an extra layer of stochasticity to reflect this effect, since whatever m-mode is set up in the disc may affect the subsequent fragmentation radius.

Despite the relatively short timescales of our simulations (since we did not make use of sink particles), we can see that orbital properties of fragments are drastically altered by interactions with each other. Since disc fragmentation forms objects with initially low eccentricities ($e < 0.1$), it is generally accepted that measurements of eccentricity as a function of orbital distance will constrain the formation mechanism of giant planets and brown dwarfs (Vorobyov, 2013), with high eccentricity being caused by dynamical scattering. However, our results in Figure 5.6 possibly suggest that these high eccentricity orbits could be formed at very early times, during the gas phase of the disc, and as such eccentricity measurements of brown dwarfs and giant planets may not, necessarily, constrain their formation mechanism.

We have shown that the initial orbital inclination of our fragments is reduced by a factor of ~ 100 over the duration of the simulation (Figure 5.7), despite the significant dynamical interactions many of the fragments experience. This suggests that although dynamical interactions certainly can create highly inclined orbits, doing so while the gas disc is present may be much more difficult. However, our simulations are of discs in isolation. Inclination and eccentricity may be excited by environment, such as a stellar companion, or location within a cluster environment (Forgan et al., 2015). Since current GI population synthesis models

do not include eccentricity or inclination information, we have provided several Gaussian fits in Figure 5.7 from our SPH simulation data. Despite our small sample size, we hope these plots will be useful in further development of GI population synthesis models.

Of the 41 clumps that were detected in this simulation, 7 were tidally destroyed ($\sim$ 20%), and 2 have orbits with eccentricity approaching unity ($e \sim 0.75$), which suggests that they are on their way to being ejected ($\sim$ 5%). If these clumps *are* ultimately ejected, then gravitational instability could, perhaps, also contribute to the population of free-floating planets (Forgan et al., 2015; Rice et al., 2003c). We have demonstrated that the orbital and structural evolution of neighbouring fragments are linked; we recommend, therefore, that any future population synthesis models are able to account for this.

5.4 Acknowledgements

We would like to thank Daniel Price for his publicly available SPH plotting code `SPLASH` (Price, 2007), which we used to produce Figures 5.9, 5.15 and 5.16. KR gratefully acknowledges support from STFC grant ST/M001229/1. DF gratefully acknowledges support from the ECOGAL project, grant agreement 291227, funded by the European Research Council under ERC-2011-ADG. The research leading to these results also received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 313014 (ETAEARTH).

See now the power of truth; the same experiment which at first glance seemed to show one thing, when more carefully examined, assures us of the contrary.

Galileo Galilei, Dialogues Concerning Two New Sciences

6

Directly Observing Self-Gravitating Spiral Waves

Work presented in this Chapter has been published in *Monthly Notices of the Royal Astronomical Society* (Hall et al., 2016). CH is the first author, and the co-authors are Duncan Forgan, Ken Rice, Tim Harries, Pamela Klaassen and Beth Biller.

6.1 Introduction

It is widely accepted that low-mass stars form through the collapse of cold, dense molecular cloud cores (McKee & Ostriker, 2007; Terebey et al., 1984). These cores will typically contain some amount of angular momentum, meaning that all the mass cannot fall directly onto the central protostar; some must first pass through a protostellar accretion disc. In these discs, molecular viscosity alone does not exert large enough torques to redistribute angular momentum out to large radii, allowing mass to accrete onto the central protostar. However, instabilities that develop into the turbulent regime can produce considerable torques that can then drive mass transport.

If these discs are sufficiently massive then self-gravity could be significant, and the gravitational instability (GI) could be the main angular momentum transport mechanism (Laughlin & Bodenheimer, 1994; Toomre, 1964) during these early times. If GI is significant in these discs, then we would expect there to be non-axisymmetric structures, typically spiral density waves.

Discs around very young stars are, however, heavily embedded in their cloud cores, making them difficult to observe at optical wavelengths (Dunham et al., 2014). High resolution interferometric observations, in radio or sub-mm, are therefore required to resolve the disc. Currently, however, observations of this wavelength with a high enough resolution to resolve spiral arms are rare. When this work was being conducted, there were no examples of spiral structure imaged at mm wavelengths, leaving unanswered our questions about the nature of spirals in imaged protostellar discs, namely, are they merely surface features? At the

time of writing of this thesis, an example of spiral structure, imaged at mm wavelengths, that *may* be due to disc self-gravity, has become apparent. Elias 2-27, located at a distance of $\sim$ 139 pc in the ρ-Ophiuchus star-forming region, was imaged by Pérez et al. (2016) at 1.3mm using the Atacama Large Millimeter Array (ALMA), revealing two-armed spiral structure extending out to $R \sim 300$ au. The central star has a mass of 0.5 $M_{\odot}$, with an estimated disc mass of $0.04 \sim 0.14$ $M_{\odot}$. If the disc mass is at the upper end of this range, then the disc-to-star mass ratio is $q \sim 0.3$, which means that the disc is susceptible to self gravity. However, since the relationship between number of spiral arms, m, and q is $m \sim 1/q$, then if the disc around Elias 2-27 is self gravitating, we may expect to observe $\sim 3-4$ spiral arms, not 2. Additionally, beyond $r \sim 100$ au, discs become susceptible to fragmentation, and it does not appear as if this disc has fragmented. Therefore, if the spiral features are due to disc self-gravity, the disc will need to be in a marginally unstable state, i.e. being either externally radiated, or sustained by mass infall from an envelope.

Here we examine the parameter space of self-gravitating protostellar discs to determine the range of accretion rates, disc masses and outer radii in which extended spiral features could be detected by ALMA. Previous studies which describe simulated ALMA observations of protostellar discs have tended to focus on reproducing the specific morphology of discs, using numerical methods such as Smoothed Particle Hydrodynamics (SPH), rather than an examination of the parameter space in which they are detectable (Cossins et al., 2010; Douglas et al., 2013; Ruge et al., 2013).

Dipierro et al. (2014) and Dipierro et al. (2015) have shown, using simulated

observations from SPH simulations, that non-axisymmetric structure in self-gravitating protostellar discs is detectable at a wide range of wavelengths using ALMA. Our approach differs to that used in Dipierro et al. (2014) and Dipierro et al. (2015), in that rather than deriving the physical disc structure from numerical simulations, where an artificial cooling law has been imposed, we use a self-consistent, analytic geometry coupled with 3D Monte Carlo radiative transfer (MCRT) to generate emission maps at typical ALMA frequencies. This geometry, unlike in SPH simulations by Dipierro et al. (2014), Dipierro et al. (2015) and Lodato & Rice (2004), is intended to be more "realistic"; by which we mean more accurately represent the amount of gravitational stress, and therefore the amplitude of the spiral arms, in the disc. Note that whenever we use the term "realistic" in this Chapter, in relation to spiral structure, *we specifically mean a more accurate representation of the amount of stress, and therefore spiral amplitude, in the disc.*

We note that we do not include in our MCRT simulations the effect of dust trapping in the spiral arms, as was done by Dipierro et al. (2015). This was originally left to future work, and we revisit this issue in chapter 7.

Once we have our MCRT images, these are then used as input sky models to the ALMA simulator from the CASA software package (ver 4.3.0) to generate synthetic ALMA images. We stress that it is not our aim to match specific morphology, but rather to examine the conditions under which the over-density in spiral arms is sufficient so as to be detected. We then use this to investigate the region of parameter space (characterised by disc mass accretion rate and outer radius) in which GI-driven spiral density waves may be detectable by ALMA.

Currently, there are few, if any, observations which are strictly comparable with what we consider here. However, several transition discs have recently been observed to have non-axisymmetric structure that extends out to large radii. Bearing in mind the difficulty of finding strictly comparable samples, we take three of these transition discs as test cases, and apply our simple geometry to them. We aim to determine if these discs exist in the parameter space where self-gravity could be a feasible explanation for their spiral structure. We do not, however, generate synthetic images of these cases, since these test cases are imaged in NIR and scattered light, which is quite different from continuum mm emission.

6.2 Semi-Analytical Model

Here we use an existing 1D model to examine the parameter space of self-gravitating discs initially developed by Clarke (2009) (see also Forgan & Rice 2013c; Rice & Armitage 2009). We develop it to include 2D and 3D structure, fitting spirals of the shape typically found in Smoothed Particle Hydrodynamics (SPH) simulations (given in Figure 6.1).

6.2.1 Radial Geometry

We expect a self-gravitating protostellar system to settle into a state of quasi-steady thermal equilibrium (Gammie, 2001; Paczynski, 1978; Rice & Armitage, 2009), with a constant accretion rate $\dot{M}$ given by

$$\dot{M} = \frac{3\pi\alpha c_s^2 \Sigma}{\Omega}, \tag{6.1}$$

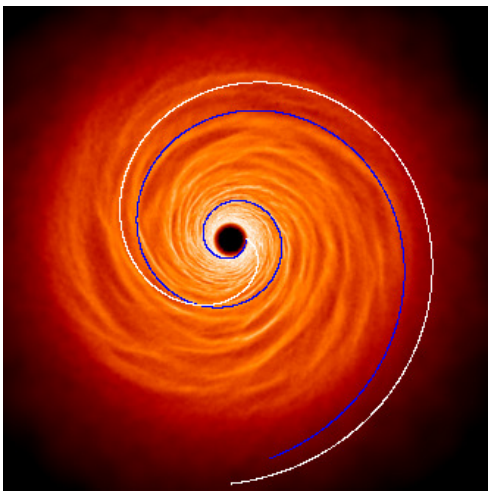


Figure 6.1: This is a simulation image of a self-gravitating disc that has reached a state of quasi-equilibrium, with parameterised cooling such that $\beta = 9$. Two spirals have been plotted. The blue line is a logarithmic spiral, the white line is an Archimedal spiral. The logarithmic spiral appears to fit the spiral arms of the disc the best, although it is fairly arbitrary since Equation (6.17) has two free parameters.

where c_s is the sound speed, Σ is the surface density, Ω is the angular frequency (since we are considering Keplerian rotation, the epicyclic frequency is simply Ω) and α is the dimensionless shear stress, composed of both hydrodynamic (i.e. Reynolds stress) and gravitational parts. Remarkably, this can be expressed both simply and analytically as (Gammie, 2001)

$$\alpha_{\rm grav} = \frac{4}{9}\frac{1}{\gamma(\gamma-1)\beta}, \tag{6.2}$$

where γ is the ratio of specific heats and β is a dimensionless constant which parameterises cooling, and is given by

$$\beta = t_c\Omega, \tag{6.3}$$

where t_c is the cooling time. The scale height H is given by

$$H = \frac{c_s}{\Omega} \tag{6.4}$$

and the midplane density is given by

$$\rho_0 = \frac{\Sigma}{2H}. \tag{6.5}$$

For a chosen $\dot{M}$ and outer radius R, we iterate our code until the surface density Σ produces the accretion rate we are attempting to match. A disc is susceptible to gravitational instability if the Toomre (1964) parameter $Q \approx 1.5-1.7$ (Durisen et al., 2007). For our purposes, we impose the following condition at all radii in the disc,

$$Q = \frac{c_s \Omega}{\pi \mathrm{G} \Sigma} = 2, \tag{6.6}$$

where G is the gravitational constant. With the chosen accretion rate and outer radius fixed, we then find Σ by guessing an overly large value and iterating downwards until $\dot{M}$ in Equation (6.1) matches our chosen $\dot{M}$. For a given value of Σ, we obtain the local sound speed by rearranging Equation (6.6). This allows the calculation of the local scale height, and hence the midplane density. An equation of state table is then used to determine the opacity κ from this density and sound speed, and then the optical depth τ is estimated as $\tau = \kappa\Sigma$ using Rosseland mean opacities from Bell & Lin (1994). The cooling rate is then (Hubeny, 1990):

$$\Lambda = \frac{8\sigma T^4}{3\tau}, \tag{6.7}$$

where T is the midplane temperature and σ is the Stefan-Boltzmann constant. If some source of external irradiation is present, this is modified to

$$\Lambda = \frac{8\sigma(T^4 - T_{\rm irr}^4)}{3\tau}, \tag{6.8}$$

where $T_{\rm irr}$ is the temperature set by the external irradiation. The cooling time t_c in both cases is then the thermal energy per unit area divided by the cooling rate:

$$t_c = \frac{1}{\Lambda}\frac{c_s^2\Sigma}{\gamma(\gamma-1)}. \tag{6.9}$$

Since in a quasi-steady state, energy dissipation in a disc is dominated by self-gravity (Lodato & Rice, 2004), and assuming local thermodynamic equilibrium, the rate per unit time, per unit area at which the kinetic energy of rotation of the disc is dissipated into heat by viscosity (the dissipation rate) is equal to the cooling rate, given by Equation 6.7. This dissipation rate is

$$\mathcal{D}(R) = \frac{1}{2}\nu_{\rm vis}\Sigma\left(R\frac{\partial\Omega}{\partial R}\right)^2, \tag{6.10}$$

where $\nu_{\rm vis}$ is kinematic viscosity, and is expressed as (Shakura & Sunyaev, 1973)

$$\nu_{\rm vis} = \frac{\alpha c_s^2}{\Omega}. \tag{6.11}$$

This allows α to absorb the uncertainties of the pseudo-viscous properties of the disc. Since in a steady state, the cooling rate matches the dissipation rate, we can equate Equations (6.7), or (6.9) and (6.10), and then use Equation (6.11) to directly determine α. Finally, since the accretion rate in a quasi-steady

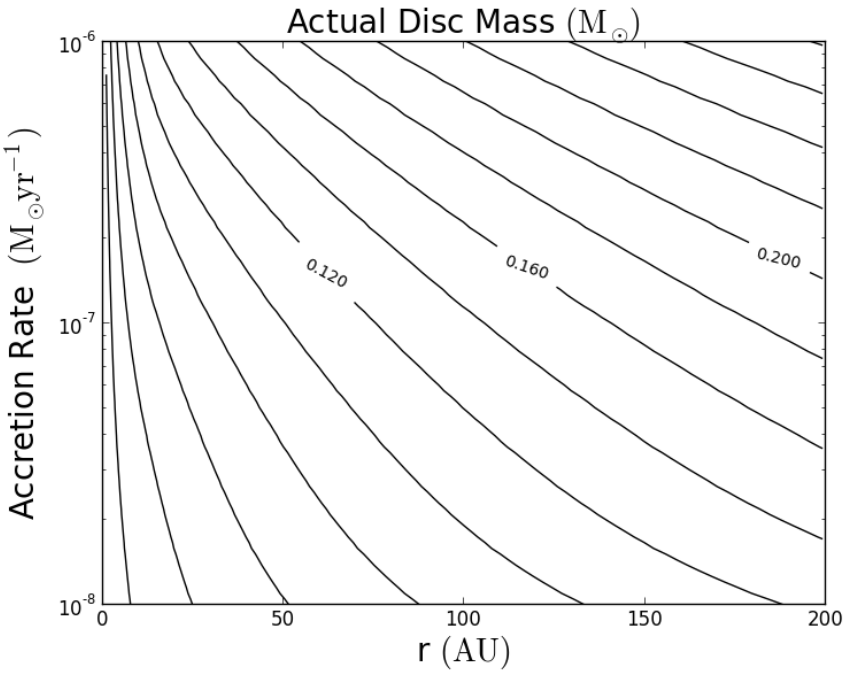


Figure 6.2: 2D contour lines of actual disc mass (around a 1 $M_\odot$ star) as a function of accretion rate and radius for self-gravitating discs with no external irradiation.

disc is given by Equation (6.1), we can check that our calculated $\dot{M}$ is within some tolerance of our imposed $\dot{M}$. If not, we reduce Σ and repeat this until $\dot{M}$ is within some tolerance of our chosen $\dot{M}$. Therefore, the surface density is determined iteratively at every radius and only needs to be integrated to determine the enclosed mass of the disc. The parameter space of our model is shown in Figure 6.2. For every accretion rate, and radius at that accretion rate, there is a corresponding disc mass.

6.2.2 3D Structure

The above method only examines the radial parameter space. To examine structures such as spiral arms, the model must be developed to manage azimuthal asymmetry. Additionally, since we later compute global 3D radiative transfer calculations, we also include a vertical density profile. We begin by defining a Cartesian grid onto which surface density as a function of radius is mapped. This produces a 2D disc which is azimuthally uniform at each radius. It has been shown there is a relationship between the amplitude of the density perturbations and the strength of the cooling such that (Cossins et al., 2009)

$$\left(\frac{\delta\Sigma}{\Sigma}\right)^2 = \frac{2}{\epsilon\beta}\frac{1}{\gamma(\gamma-1)}\left(\frac{1}{\mathcal{M}\widetilde{\mathcal{M}}}\right), \tag{6.12}$$

where ϵ is a dimensionless proportionality factor known as the heating factor, and the radial phase and Doppler-shifted phase Mach numbers are $\mathcal{M} = |v_\mathrm{p}|/c_s$ and $\widetilde{\mathcal{M}} = |\tilde{v}_\mathrm{p}|/c_s$. Cossins et al. (2009) find an empirical relationship such that

$$\frac{\langle\delta\Sigma\rangle}{\langle\Sigma\rangle} \approx \beta^{-\frac{1}{2}}. \tag{6.13}$$

Since there is a relationship between α and β given by Equation (6.2), we impose a spiral perturbation dictated by our α_grav such that (Rice et al., 2011)

$$\frac{\langle\delta\Sigma\rangle}{\langle\Sigma\rangle} \approx \alpha_\mathrm{grav}^{1/2}, \tag{6.14}$$

where $\delta\Sigma/\Sigma$ is the fractional over-density. This is imposed sinusoidally so that

$$\delta\Sigma(\phi) = -\langle\delta\Sigma\rangle \cos(m\phi), \tag{6.15}$$

where m is the azimuthal wavenumber selected by the user. For self-gravitating discs with mass ratios similar to what we will be considering, this is typically high, so we select 10. We have assumed that the maximum over-density is equal to the average over-density, with ϕ the phase difference between the azimuthal location of the spiral $\theta_{\rm spiral}$ and the azimuthal position $\theta_{\rm x,y}$ in the disc:

$$\phi = \theta_{\rm spiral} - \theta_{\rm (x,y)}. \tag{6.16}$$

The maximum over-density then occurs when the (x,y) position is coincident with the position of the spiral. The negative sign in Equation (6.15) simply forces the maxima to occur coincident with the position of the spiral. A logarithmic spiral of the form

$$\theta_{\rm spiral} = \frac{1}{b}\log\left(\frac{r}{a}\right), \tag{6.17}$$

was used, where a and b are constants (in this case $a = 13.5$ and $b = 0.38$), as this most closely matches the shape of spirals seen in self-gravitating discs from simulation data (see Figure 6.1). This is somewhat arbitrary, since tweaking parameters can give a close fit with an Archimedean spiral, which has an equation of the form

$$r = a\theta^{\frac{1}{n}}, \tag{6.18}$$

where n is a constant that determines how tightly wound the spiral is. However, it is not our intention to exactly match morphology.

To create accurate skymodels in terms of brightness to put into the ALMA simulator, the vertical density profile must be carefully considered, since the total intensity at the surface of the disc is dependent upon the amount of emission and absorption which has occurred between the surface and the midplane. We calculate the density as a function of z using the expression for density in a self-gravitating disc (see e.g. Spitzer (1942) for a full derivation)

$$\rho(z) = \rho_0 \left[\frac{1}{\cosh^2\left(\frac{z}{H_{\rm sg}}\right)} \right], \tag{6.19}$$

where the self gravitating scale height $H_{\rm sg}$ is given by

$$H_{\rm sg} = \frac{c_s^2}{\pi \mathrm{G}\Sigma}. \tag{6.20}$$

It is worth noting that H_{sg} is approximately equal to H in Equation (6.4) since

$$\frac{H_{\rm sg}}{H} = \frac{c_s \Omega}{\pi \mathrm{G}\Sigma} = Q = 2. \tag{6.21}$$

6.3 Radiative and Molecular Line Transfer Code: TORUS

The TORUS radiation transport code (Harries et al., 2004; Haworth et al., 2015; Kurosawa et al., 2004) determines radiative equilibrium in a dusty medium using

the Monte Carlo (MC) photon packet method first described by Lucy (1999). Temperatures, densities and dust properties are stored on a three-dimensional adaptive mesh, refined in such a way that large gradients in opacity are well resolved. Here we use a cylindrical adaptive mesh, in which the cells are sectors of hollow cylinders. When additional resolution is required a cell may be subdivided into four (or eight) children by splitting the radial extent and height of the cell into two, and (optionally) by splitting the azimuthal extent of the sector into two.

We use a simple thermal model for the disc geometry, and then calculate a radiative-equilibrium model using TORUS. The radiation equilibrium calculation is iterative and full details are given in the above references. Briefly, the radiation field of the protostar is modelled using N photon packets which are allowed to propagate through the grid, undergoing a random walk of scatterings, or absorptions and re-emissions, until they escape the computational domain, at which point estimates are made of the absorption rate in each cell. The dust temperatures in the grid are then calculated on a cell-by-cell basis by assuming radiative equilibrium, and the next iteration of the photon loop is performed (using the updated dust temperatures). Once the temperatures have converged it is possible to calculate spectral energy distributions and continuum images for arbitrary viewing angles using the MC method. The TORUS code has been extensively benchmarked and shows good agreement with other independently developed radiative transfer codes (Pinte et al., 2009).

For our main radiative transfer results (in Section 6.5.3) we assume typical values for a pre-main-sequence star, with central source mass of $M_* = \mathrm{M}_\odot$, $R_* = 2.325$ $R_\odot$ and $T_{\rm eff} = 4350$ K. The dust in our model consists of Draine & Lee (1984)

silicates, with a grain size distribution of

$$n(a) \propto a^{-q} \qquad \text{for} \qquad a_{\rm min} < a < a_{\rm max}, \tag{6.22}$$

where $a_{\rm min}$ and $a_{\rm max}$ are the minimum and maximum size of the dust grains (0.1 μm and 2000 μm respectively), and the power-law exponent q is taken to be $q_{\rm ism} = 3.5$ (Mathis et al., 1977). The dust density is 1% of the gas density.

6.4 The ALMA Simulator

The emission maps generated by TORUS are used as inputs to the ALMA simulator built into CASA (ver 4.3) (McMullin et al., 2007). Disc sizes and fluxes are scaled to a distance of $\sim$ 140 pc (i.e. in Taurus), and we show a set of comparison images at $\sim$ 50 pc (i.e. the T-Tauri star TW Hydrae, located in Hydra).

Multiple simulations were conducted varying the array size, and therefore imaging resolution and sensitivity, to ensure the optimum balance between resolution and sensitivity. Built-in noise sources such as atmospheric transmissions were included in the simulations run at varying precipitable water vapour (PWV) levels. Typical PWV levels appropriate for observing in the different ALMA bands were used, specifically: 2.784, 1.262, 1.262 and 0.472 mm for simulated observations at 220, 345, 460 and 680 GHz, respectively. These are consistent with those used, for instance, in Dipierro et al. (2014).

6.5 Results

Our results section is broken into four parts. We begin with the basic results from our analytic geometry (Section 6.5.1), showing that for a given central star mass and disc outer radius, there are regions of parameter space in which the disc cannot exist in a quasi-steady, self-gravitating state without fragmenting. We then show that such discs may exist in these regions of parameter space if they are irradiated with some external source, but a moderate amount of irradiation will remove spiral features.

In Section 6.5.2, we compare the synthetic ALMA image results of our analytical model to image results which used SPH geometries. We show that the assumptions made in SPH simulations probably cause larger contrast in the inner regions of discs. We do, however, reproduce the basic results of Dipierro et al. (2014) using a constant β approach.

In Section 6.5.3, we present results showing the conditions required to directly observe self-gravitating spiral structures with ALMA. We entered a range of accretion rates, and found that even at the maximum accretion rate a 100 AU disc can sustain without fragmenting, non-axisymmetry due to disc self-gravity is just detectable at 680 GHz at a distance of 140 pc (i.e. in Taurus). At 50 pc (i.e. TW Hydrae), it is significantly easier to detect spiral structure, and the features are discernible at 220 GHz.

Finally, we make only parameter space comparisons with three observed systems in Section 6.5.4. We show that for all three systems, it seems unlikely that spiral features which have recently been imaged in the disc are due to disc self-gravity,

unless the disc mass has been significantly underestimated.

6.5.1 Basic Model

We find that for a given radius and host star mass, with no external irradiation considered, there is a maximum accretion rate that any self-gravitating, quasi-steady disc can sustain. Above this accretion rate, the disc begins to truncate as the outer parts become susceptible to fragmentation (Forgan & Rice, 2011; Rafikov, 2005). This is illustrated in the top panel of Figure 6.3, which shows logarithmic surface density plots of three discs with increasing accretion rate from left to right. Note that to decrease the disc radius to a quarter of its original size only requires an increase in accretion rate of around 30%.

When more mass is added to the disc, the sound speed (and subsequently temperature from our equation of state) increases since in our model it is set by Equation (6.6). In the cool outer parts of the disc, the Rosseland mean opacity is related to temperature by $\kappa \propto T^2$ (Whitworth et al., 2010). Therefore our cooling rate now has a dependence $\Lambda \propto T^2$. This increased local cooling rate causes a decrease in local cooling time. In order to maintain the disc in a quasi-steady state, this local radiative cooling must be balanced by viscous shock heating from the spiral arms, therefore the local $\alpha_{\rm grav}$ increases to redress the balance in the disc.

However, this quasi-steady, self-gravitating torque saturates at around $\alpha \sim 0.1$ (Gammie, 2001; Rice et al., 2005b) and we expect the disc to fragment, producing bound objects. Since this region of parameter space is not what we are interested

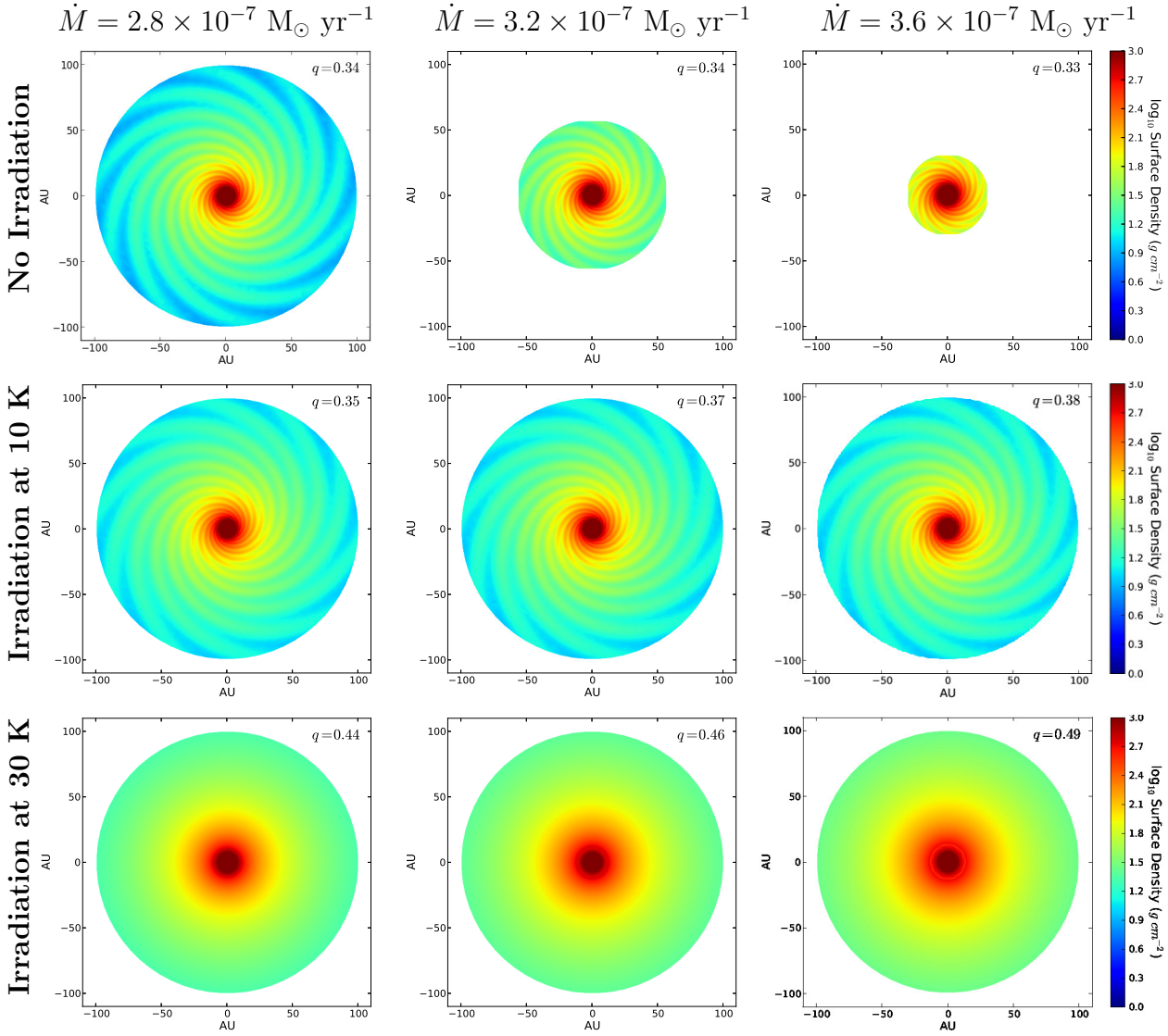


Figure 6.3: Logarithmic surface density maps of discs with accretion rate increasing from left to right. Top row: For a 1 M$_\odot$ central star and no external irradiation, the maximum accretion rate a 100 AU disc can sustain without the outer regions fragmenting is $\dot{M} = 2.8 \times 10^{-7}M_\odot$ yr^{-1}. Any higher, and the disc truncates in the outer regions as it becomes unstable to collapse and begins to fragment. Middle row: same three accretion rates, but with external irradiation at 10K added. Although the discs look similar, they vary very slightly in mass. The irradiation is just enough to prevent truncation, without removing spiral features. Bottom row: same discs but with irradiation at 30K added. These discs are also prevented from fragmentation in the outer regions, however the irradiation is also sufficient to remove spiral features completely.

in, we set the surface density here to 0.

However, a small amount of irradiation, say at 10K, can change the surface density profile of a disc. How this changes depends upon the accretion rate of the disc, and therefore how close $\dot{M}$ was initially to the fragmentation/truncation boundary in the absence of irradiation.

The middle panels of Figure 6.3 show discs with an accretion rate that was initially at the fragmentation boundary before irradiation at 10K was added. In this case, the irradiation prevents truncation whilst preserving spiral structure. The disc mass varies very slightly, and this is consistent with previous examinations of disc mass under irradiation (Forgan & Rice, 2013a) since the accretion rate is only changing slightly, rather than by an order of magnitude. We therefore find that a small amount of external irradiation can prevent fragmentation, whilst preserving the spiral features in the disc.

On the other hand, we do find that spiral features in a self-gravitating disc of any accretion rate are erased by a moderate amount of external irradiation, say at 30K. This is demonstrated by the bottom panels in Figure 6.3, which shows that for a disc with a radius of 100 AU, irradiation at a temperature of 30K can erase non-axisymmetric structure from the disc. It is unsurprising that applying sufficient external irradiation to our systems wipes out spiral structure, as external irradiation has been found to have a stabilising effect on marginally unstable discs (Forgan & Rice, 2013a; Kratter & Murray-Clay, 2011; Rice et al., 2011). When external irradiation of temperature $T_{\rm irr}$ is present, the local cooling rate in the disc is reduced and is given by Equation (6.8). This decrease in cooling rate causes

the cooling time $t_c = \beta/\Omega$ to increase, which also causes local β to increase. If the disc is in thermal equilibrium, we then parameterise the viscous shock heating required to balance this cooling as $\alpha_{\rm grav}$ given by Equation (6.2), where $\alpha_{\rm grav}$ can be thought of as an effective gravitational stress. It is easy to see that for an increased β, a smaller α is required to redress the balance. The α is then able to stay below the torque saturation limit $\alpha_{\rm crit} \sim 0.1$ (Gammie, 2001; Rice et al., 2005b), and the disc is able to stay in a quasi-steady self-gravitating state out to larger radii. Although in this manner, the disc is able to stay in a quasi-steady, self-gravitating state, the strength of the spiral amplitudes decrease. In our calculations, the α is composed of two parts,

$$\alpha = \alpha_{\rm grav} + \alpha_{\rm visc}, \tag{6.23}$$

where $\alpha_{\rm visc}$ is the viscous component of the stress due to magnetorotational instability (MRI), which we set to $\alpha_{\rm visc} = 0.01$ (see e.g. Kratter et al. 2008). Since the perturbation strength of the spiral amplitudes is given by

$$\frac{\langle \delta\Sigma \rangle}{\langle \Sigma \rangle} = \alpha_{\rm grav}^{\frac{1}{2}}, \tag{6.24}$$

when $\alpha \to 0$ then $\frac{\langle \delta\Sigma \rangle}{\langle \Sigma \rangle} \to 0$. This happens when the midplane temperature T in Equation (6.8) tends to $T_{\rm irr}^4$. In this case, cooling rate $\Lambda \to 0$, so cooling time $t_c \to \infty$ and $\beta \to \infty$. This means $\alpha_{\rm grav} \to 0$ by Equation (6.2), and $\frac{\langle \delta\Sigma \rangle}{\langle \Sigma \rangle} \to 0$. In this limit, cooling is balanced by irradiation and $\alpha_{\rm visc}$, so $\alpha_{\rm grav}$ is no longer needed for thermal equilibrium. Essentially, increasing temperature provides extra pressure support against gravitational collapse. In our geometry, however, we do not

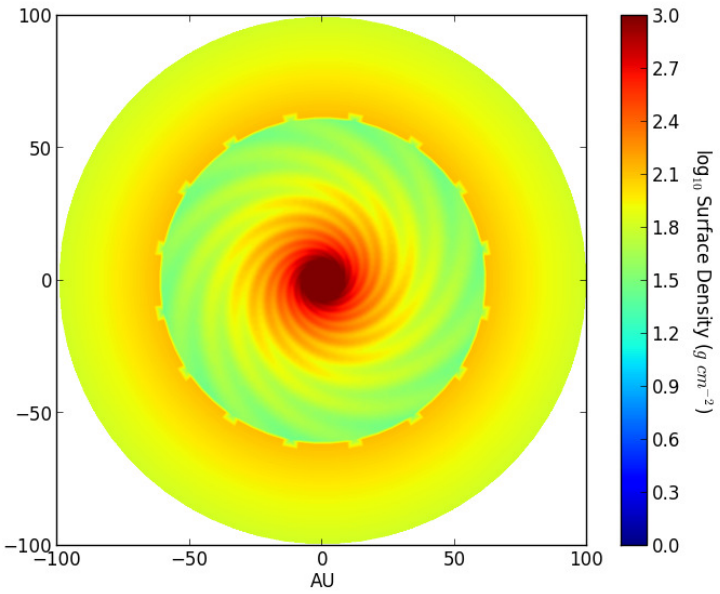


Figure 6.4: Disc with $\dot{M} = 2.0 \times 10^{-7}$ $M_\odot$ yr^{-1}, 33% lower than the maximum it can sustain without truncating. It is irradiated at 10 K, and this small amount of external irradiation has altered the surface density structure of the outer parts of the disc, as the equilibrium disc structure is now more massive for a given accretion rate and radius. The "spikey" features on the boundary between the outer and inner parts of the disc are a numerical artefact due to a small change in surface density from the resolution limit of the grid, and is more pronounced in log space.

consider infalling mass from a nascent cloud. This could potentially reverse our result of small amounts of irradiation halting fragmentation, as this infalling mass causes a positive rate of change of local mass, decreasing the Jeans mass and potentially encouraging fragmentation, provided that Q remains constant.

Lowering the accretion rate (and therefore disc mass), changes the effect of a given amount of external irradiation. Figure 6.4 illustrates this. It is a disc with an accretion rate of $\dot{M} = 2.0 \times 10^{-7}$ $M_\odot$ yr^{-1}, approximately a third lower than in Figure 6.3. In this case (Figure 6.4) external irradiation has added additional

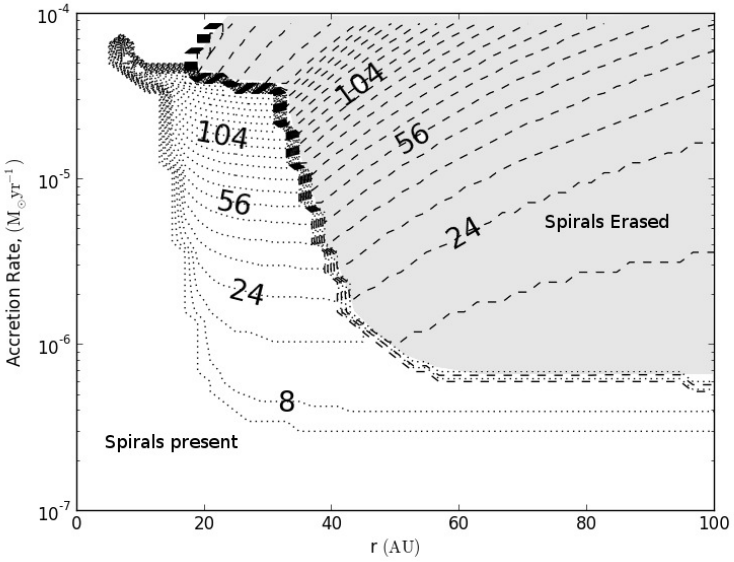


Figure 6.5: Contour plot showing the minimum temperature in (K) of external irradiation required to halt fragmentation as a function of accretion rate and radius. Shaded region with dashed contours shows where both fragmentation is suppressed and spiral features are erased. The unshaded region with dotted contours shows where added irradiation prevents fragmentation but preserves spiral structure.

mass to the outer parts of the disc, as the equilibrium disc structure is now more massive for a given accretion rate and radius, although in this case it does not meet the criteria required for it to fragment (which requires that the local $\alpha \gtrsim 0.1$). In this disc, the external irradiation is maintaining $Q \sim 2$ with little dissipation of the gravitational instability. In reality, we should expect Q to increase beyond the marginal limit of self-gravity in the outer parts of such discs.

We should bear in mind that this is a simple model in which no infall is considered. Since, in general, adding mass to the outer regions of a self-gravitating disc encourages fragmentation (Forgan & Rice, 2012; Kratter & Murray-Clay, 2011; Kratter et al., 2010a; Vorobyov & Basu, 2010), it seems that the Jeans criterion

for fragmentation may be satisfied at relatively high accretion rates (of order $\sim 10^{-7}$ $M_\odot$ yr^{-1}) in the presence of sufficiently small (~ 10 K) irradiation.

Figure 6.5 is a contour plot of the amount of external irradiation required to prevent fragmentation of the disc as a function of accretion rate and radius. Below $\sim 3.2 \times 10^{-7}$ $M_\odot$ yr^{-1}, in our local viscous approximation, any disc will be able to regulate itself against collapse. Above this, differing amounts of irradiation can either prevent fragmentation and totally remove spiral features from the disc, or prevent fragmentation whilst allowing spiral features to exist. As accretion rate increases, higher temperatures are required to prevent the disc from fragmenting; additionally, beyond 60 AU the determining factor in whether spirals are present or not is temperature, rather than a combination of temperature and accretion rate.

6.5.2 Comparison with imposed constant β

In our model, unlike in SPH simulations by Dipierro et al. (2014) and Lodato & Rice (2004), we do not artificially impose a constant gravitational $\alpha_{\rm grav}$ by imposing a constant β. Doing so implies that the spiral amplitude strength $\langle \delta\Sigma \rangle / \langle \Sigma \rangle \approx \alpha_{\rm grav}^{1/2}$ is constant at all radii. In our model we take a more realistic approach and allow α to vary, requiring that our disc remain marginally unstable to non-axisymmetric perturbations (i.e. $Q = 2$), and that the cooling rate at each radius depends on the local conditions. In such a quasi-steady, self-gravitating disc we expect α to increase with increasing radius (Rice & Armitage, 2009).

The basic consequence of this is that the assumption of a constant β means that

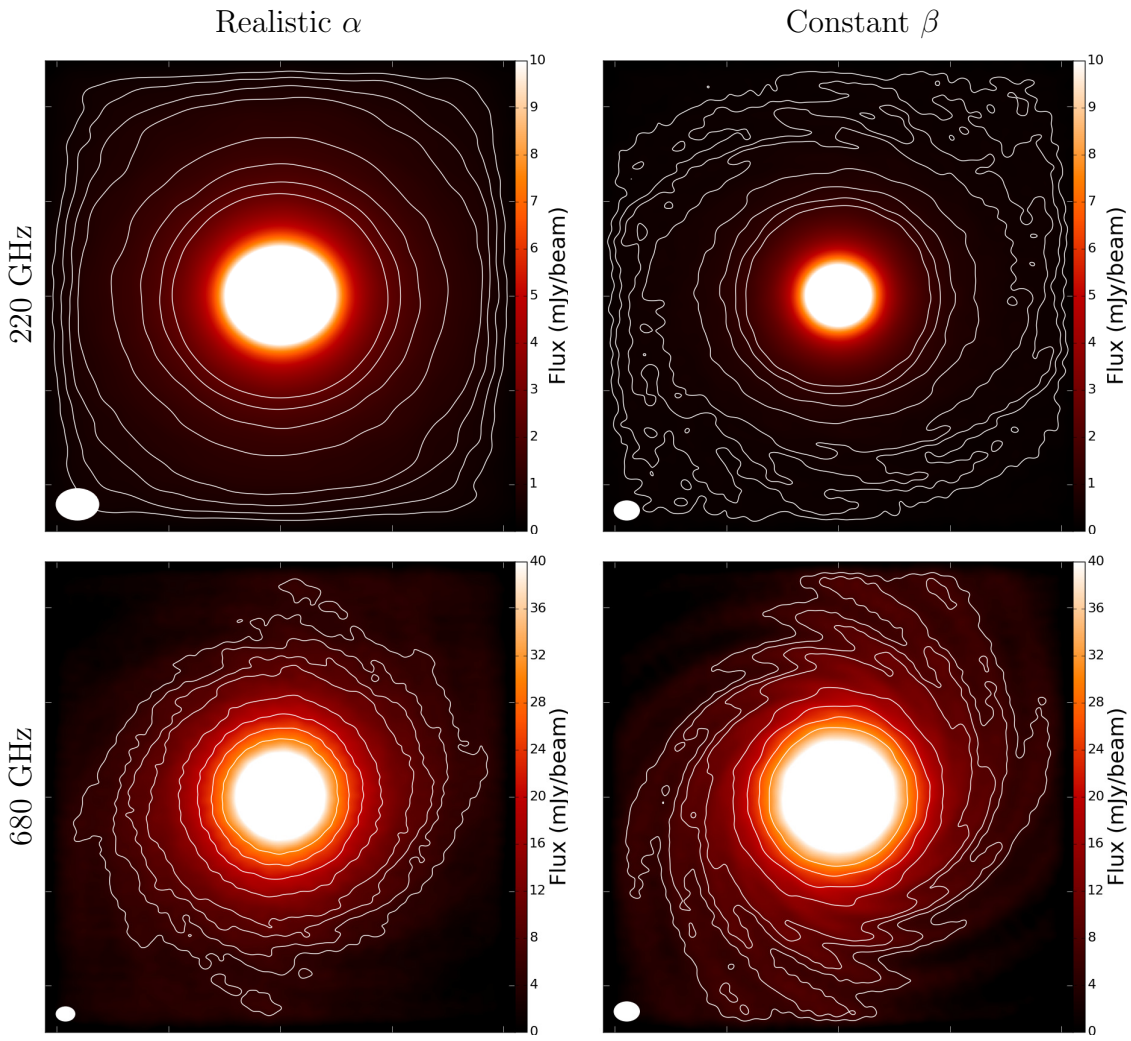


Figure 6.6: Synthesised ALMA images for two $R = 100$ AU discs with different geometries. Both discs have a central star mass of 1 $M_\odot$ and central star radius 1 $R_\odot$, with $T_{\rm source} = 6000K$, imaged at a distance of 50 pc for clarity. Beam size which gives the best compromise resolution and sensitivity is selected for each disc, details are given in Table 6.1. Left column has realistic α, whereas right column has in imposed constant β (and therefore α) value. Contours are at 3, 4, 5, 6, 9, 12, 15 & 18 $\times$ the RMS in each image. Each image has been scaled to best show the spiral features (if present). Disc mass is the same for both ($q = 0.21$), $\dot{M} = 5 \times 10^{-8}$ $M_\odot$ yr^{-1} for realistic α and $\dot{M} = 2.8 \times 10^{-7}$ $M_\odot$ yr^{-1} for constant β - and, hence, constant α. In both cases, the spiral arms are more clear for $\beta = 6$.

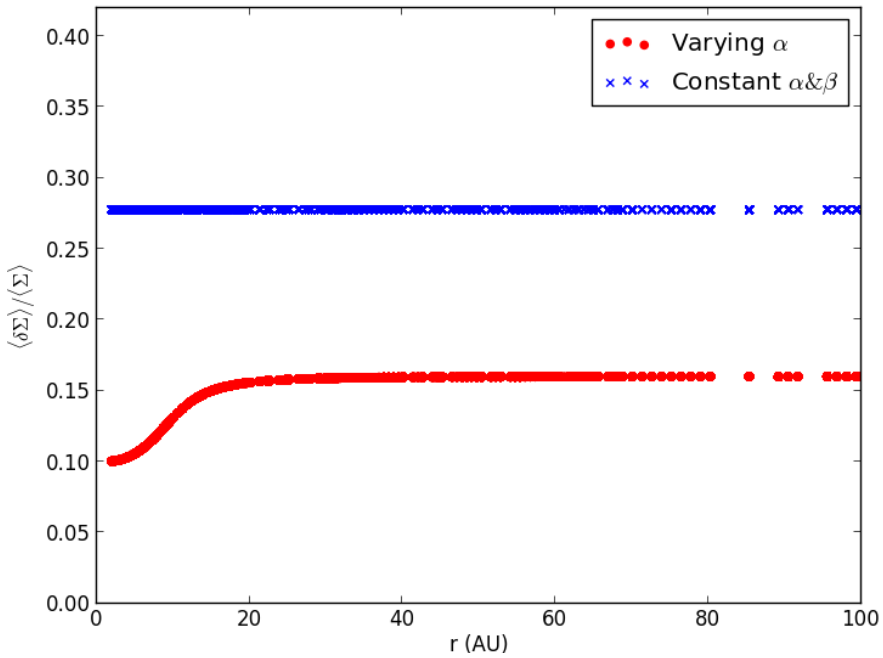


Figure 6.7: Spiral amplitude strength vs radius for the two discs. Red dots are "realistic" α and blue crosses are fixed $\alpha(\beta)$.

there will clearly be regions where the α values will be quite different to what would be the case if more realistic assumptions were used. In particular, the α values would probably be significantly larger in the inner regions, which will cause a greater contrast than would be the case were more realistic assumptions used. Figure 6.6 compares two systems, both with the same total mass, but one determining α realistically (left hand panel) and the other assuming that β (and hence α) is constant, fixed at $\beta = 6$. The reason we choose the same disc mass, rather than the same accretion rate, is that the total disc flux $F_{\rm D}$ is related to total disc mass $M_{\rm D}$ by

$$F_D = \frac{2\kappa(\nu)\mathrm{k_B}\nu^2 T_{\rm dust} M_{\rm disc}}{D^2 c^2}, \tag{6.25}$$

where $\kappa(\nu)$ is opacity, k_B is the Boltzmann constant, c is the speed of light and D is the distance to the object. The disc mass is the same in both cases ($q = 0.2$), but holding β fixed alters the geometry so the equilibrium accretion rate for a given disc mass differs. Therefore $\dot{M} = 5 \times 10^{-8}$ $M_\odot$ yr^{-1} for realistic α and $\dot{M} = 2.8 \times 10^{-7}$ $M_\odot$ yr^{-1} for constant β yield the same total disc mass.

In the fixed β (and therefore fixed α) scenario, the strength of the spiral amplitudes are larger throughout the disc and so they are easily detectable to 3σ confidence level even at $\nu = 220$ GHz. At $\nu = 680$ GHz, the interarm regions are a low enough temperature to broach the Wien limit, thus increasing the contrast ratio between the spiral arms and the inter-arm regions.

The first thing that we should stress is that using our semi-analytic model with fixed beta, we are able to reproduce the basic results of Dipierro et al. (2014). We find that extended structure is detectable to 3 $\times$ the RMS noise, and that the fluxes agree to the same order of magnitude in both the realistic α case and imposed constant β (hence α) case, giving us some confidence that our general method is reasonable.

We have, however, assumed our dust grains are dynamically well coupled to the gas such that $\rho_{\rm dust}$ is linearly proportional to $\rho_{\rm gas}$. Dipierro et al. (2015) showed that solid particle trapping occurs in the spiral arms in self-gravitating discs, however this predominantly occurs for particles with sizes of a centimetre or more (Rice et al., 2004). Therefore, these overdensities are best probed at frequencies that are not considered here ($\sim$ 10 GHz), and so should not change our results.

However, what Figure 6.7 also shows is that in the fixed β scenario, the strength

of the spiral amplitudes are larger throughout the disc than in the realistic α case. At $\nu = 680$ GHz, this translates to more clearly discernible spiral structure throughout the disc. We could, of course, choose a larger β value, or increase the mass in the realistic alpha case, so that there were at least regions where the amplitudes were comparable. However, it is clear that even in such cases, the constant β assumption would produce unrealistic amplitudes in the inner parts of these discs.

For the rest of this work we allow the effective gravitational stress $\alpha_{\rm grav}$ to vary as it would do in a "realistic" disc.

6.5.3 ALMA Images

In this section we present results illustrating the conditions under which we may be able to directly observe self-gravitating spiral density structures using ALMA. We assume that all discs have outer radii of 100 AU and follow the procedure described in Section 6.2 to determine the amplitude of the spiral density waves, the continuum emission from the disc, and what we would then expect to be observed by ALMA.

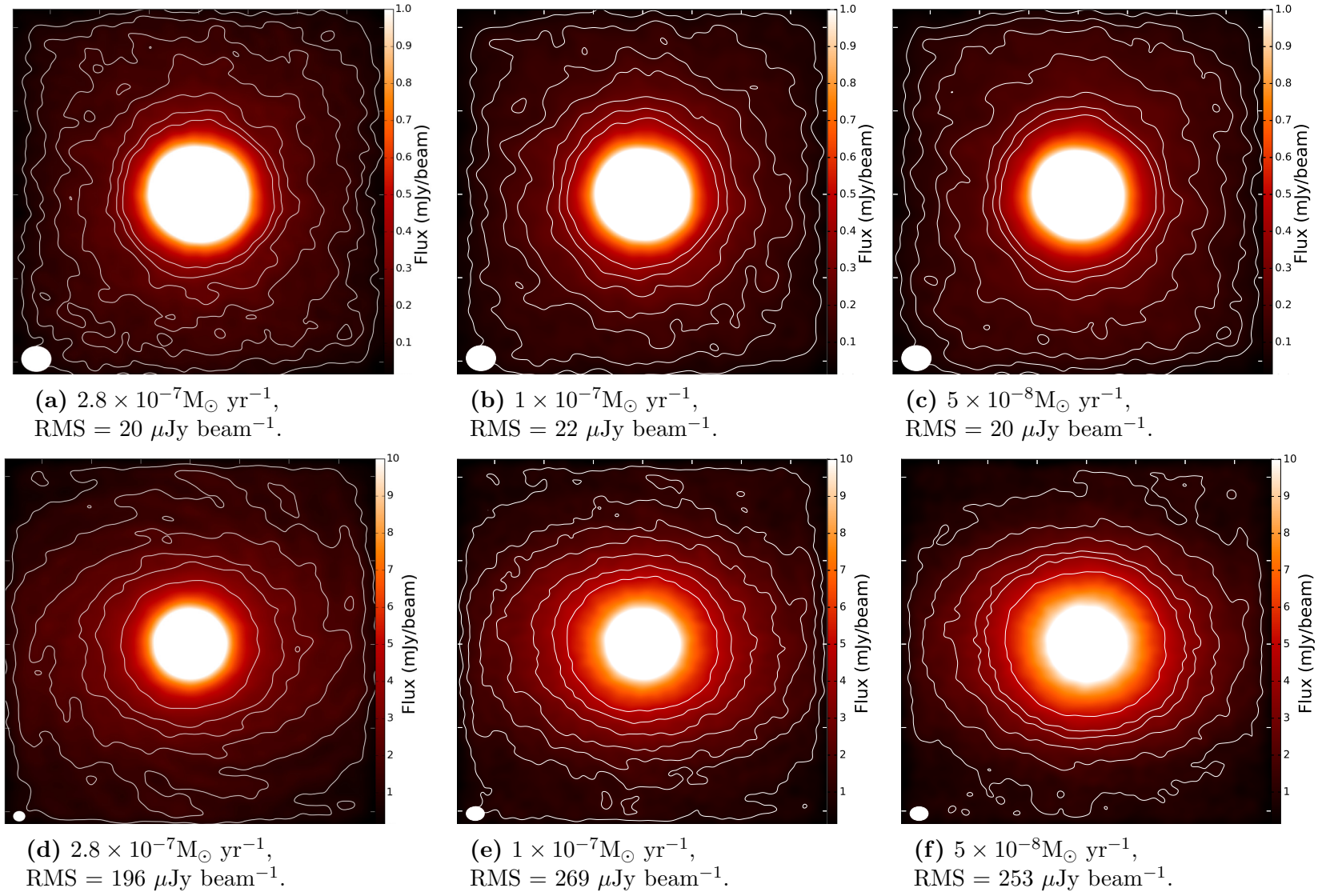
Figure 6.8 shows six synthetic ALMA images of three 100AU discs around a 1.0 $M_\odot$, 2.325 $R_\odot$ star, with a surface temperature of $T = 4350$ K, and with three different accretion rates. All discs are imaged at a distance of 140 pc. Each disc is observed at 220 GHz (top row) and 680 GHz (bottom row). Observing parameters are given in Table 6.1, and physical parameters are given in Table 6.2. The observing frequencies correspond to Band 6 of ALMA (220 GHz) and

Band 9 of ALMA (680 GHz), with resolution increasing with increasing ν.

From left to right, the accretion rates are $\dot{M} = 2.8{\times}10^{-7}$ $M_\odot$ yr^{-1}, $\dot{M} = 1.0{\times}10^{-7}$ $M_\odot$ yr^{-1} and $\dot{M} = 5.0 \times 10^{-8}$ $M_\odot$ yr^{-1}. Figures 6.8a and 6.8d depict a disc with the maximum accretion rate $\dot{M} = 2.8 \times 10^{-7}$ $M_\odot$ yr^{-1} a disc of outer radius $R = 100$ AU, with no external irradiation, can sustain without fragmenting. Non-axisymmetric structure is only visible at the higher frequency of 680 GHz. We wish to stress at this point that this is the absolute maximum accretion rate that this disc, with this particular set of parameters and no external irradiation, can sustain.

For a sufficiently high accretion rate, and therefore disc-to-star mass ratio q, we reproduce the results of Dipierro et al. (2014) of increasing contrast with increasing frequency, as the Planck law in the interarm regions falls into the Wien limit due to the low temperature.

Additionally, Figure 6.8 shows that as accretion rate (and therefore disc mass) is decreased, the central part of the disc remains detectable to at least the 3σ level, but the spiral structure is simply not detectable with ALMA at 220 GHz (or longer wavelengths). Non-axisymmetry is, however, noticeable at higher frequency (shorter wavelengths) since the lower temperatures in the inter-arm regions means the Planck law is in the Wien limit for these frequencies, reducing the intensity of emission.



(a) $2.8 \times 10^{-7} \mathrm{M}_\odot$ yr^{-1}, RMS = 20 μJy beam^{-1}.

(b) $1 \times 10^{-7} \mathrm{M}_\odot$ yr^{-1}, RMS = 22 μJy beam^{-1}.

(c) $5 \times 10^{-8} \mathrm{M}_\odot$ yr^{-1}, RMS = 20 μJy beam^{-1}.

(d) $2.8 \times 10^{-7} \mathrm{M}_\odot$ yr^{-1}, RMS = 196 μJy beam^{-1}.

(e) $1 \times 10^{-7} \mathrm{M}_\odot$ yr^{-1}, RMS = 269 μJy beam^{-1}.

(f) $5 \times 10^{-8} \mathrm{M}_\odot$ yr^{-1}, RMS = 253 μJy beam^{-1}.

Figure 6.8: Synthesised ALMA images for $R = 100$ AU discs with accretion rate decreasing from left to right. Top row is at 220 GHz, bottom row is at 680 GHz. All discs are imaged at 140 pc. Left column is the maximum accretion rate it can sustain without fragmenting, $\dot{M} = 2.8 \times 10^{-7} \mathrm{M}_\odot$ yr^{-1}. Contours are at multiples of the RMS given in Table 6.1. Beam size is in bottom left corner, beam details are given in Table 6.1. Geometry details are given in Table 6.2. Below $\dot{M} = 5.0 \times 10^{-8} \mathrm{M}_\odot$ yr^{-1}, at 220 GHz flux from the disc is low enough that thermal noise dominates, so the central region (inner 20 AU) is detectable, but not the extended non-axisymmetric structure. At 680 GHz, the outer regions of the disc display some non-axisymmetric structure, but the spiral arms are not clearly defined.

Figure	ν (GHz)	Distance (pc)	Beam size (asec)	RMS (μJy beam^{-1})	Contours ($\times$ RMS)	Integration time (s)	PWV (mm)
6.6	220	50	0.23 $\times$ 0.17	141	3,4,5,6,9,12,18	1800	2.784
	220	50	0.13 $\times$ 0.10	82	3,4,5,6,9,12,18	1800	2.784
	680	50	0.13 $\times$ 0.11	1674	3,4,5,6,9,12,18	7200	0.472
	680	50	0.13 $\times$ 0.10	1732	3,4,5,6,9,12,18	1800	0.472
6.8a	220	140	0.0706 $\times$ 0.0601	20	3,5,7,9,12,15,18	7200	2.784
6.8b	220	140	0.0706 $\times$ 0.0601	22	3,5,7,9,12,15,18	7200	2.784
6.8c	220	140	0.0706 $\times$ 0.0601	20	3,5,7,9,12,15,18	7200	2.784
6.8d	680	140	0.0435 $\times$ 0.0331	196	4,7,10,13,15,20	7200	0.472
6.8e	680	140	0.0435 $\times$ 0.0331	269	3,5,7,9,12,15,18	7200	0.472
6.8f	680	140	0.0435 $\times$ 0.0331	253	3,5,7,9,12,15,18	7200	0.472
6.9	220	140	0.0706 $\times$ 0.0601	20	3,5,7,9,12,15,18	7200	2.784
	220	50	0.135 $\times$ 0.102	60	4,7,10,13,15,18	7200	2.784
	680	140	0.0435 $\times$ 0.0331	196	4,7,10,13,15,20	7200	0.472
	680	50	0.0732 $\times$ 0.0539	720	4,7,10,13,15,20	7200	0.472

Table 6.1: Image details for figures shown in this work. We detail frequency of the synthesised observation, whether a realistic or constant α was used in the image, the accretion rate, disc mass, the size of the beam, the noise of the image, the integration time used and precipitable water vapour (PWV) value.

The images in Figure 6.8 show the difficulty faced when determining the presence of GI in a protostellar disc. Since a distance of 140 pc corresponds to a typical star forming region (Taurus), these are not the most optimistic results. Figure 6.9 shows the same disc, with an accretion rate of $\dot{M} = 2.8 \times 10^{-7}$ $M_{\odot}$ yr^{-1}, imaged at 140 pc (220 GHz and 680 GHz), and 50 pc (at 220 GHz and 680 GHz). At a closer distance, it is significantly easier to detect spiral structure in the disc, as shown in the images at 50 pc. At a distance of 50 pc, the spiral structure in the outer part of the disc is also detectable at 220 GHz, whereas it is not detectable at 140 pc.

It becomes more difficult to detect spiral structure with decreasing accretion rate and disc mass. Since the strength of our spiral amplitudes are determined by $\delta\Sigma/\Sigma \approx \alpha^{1/2}$, when the accretion rate is lowered, so are the spiral amplitudes. The conclusion we can draw from this examination of parameter space is that in quasi-steady, self-gravitating discs, for any given disc radius and host star mass, there exists a narrow range of accretion rates in which the outer part of the disc does not begin to fragment, but for which the spiral structure is detectable. Additionally, even if the disc is within this parameter space, the distance to the object and the frequency of the observations may also determine the likelihood of spiral structure being detected.

With this in mind, it is prudent to caution against diagnosing directly imaged non-axisymmetric structure as due to disc self-gravity, unless the disc in question is sufficiently massive. That our discs of lower disc mass/accretion rate fail to produce detectable spiral features appears to conflict with the result found by Dipierro et al. (2014). However, as previously mentioned, our model allows

Figure	α type	M_* ($M_\odot$)	$\dot{M}$ ($M_\odot$ yr^{-1})	q (M_D/M_*)
6.6	Realistic	1	5×10^{-8}	0.21
	Constant	1	2.8×10^{-7}	0.21
	Realistic	1	5×10^{-8}	0.21
	Constant	1	2.8×10^{-7}	0.21
6.8a	Realistic	1	2.8×10^{-7}	0.34
6.8b	”	1	1.0×10^{-7}	0.25
6.8c	”	1	5.0×10^{-8}	0.21
6.8d	”	1	2.8×10^{-7}	0.34
6.8e	”	1	1.0×10^{-7}	0.25
6.8f	”	1	5.0×10^{-8}	0.21
N/A				
MWC 758	Realistic	2	2.0×10^{-7}	0.25
SAO 206462	”	1.7	5.37×10^{-9}	0.1
HD 142527	”	2	6.92×10^{-8}	0.75

Table 6.2: Physical parameters of the discs used to create synthetic images in this work. Columns are figure number, whether a realistic or constant α was used, host star mass, accretion rate and disc-to-star mass ratio

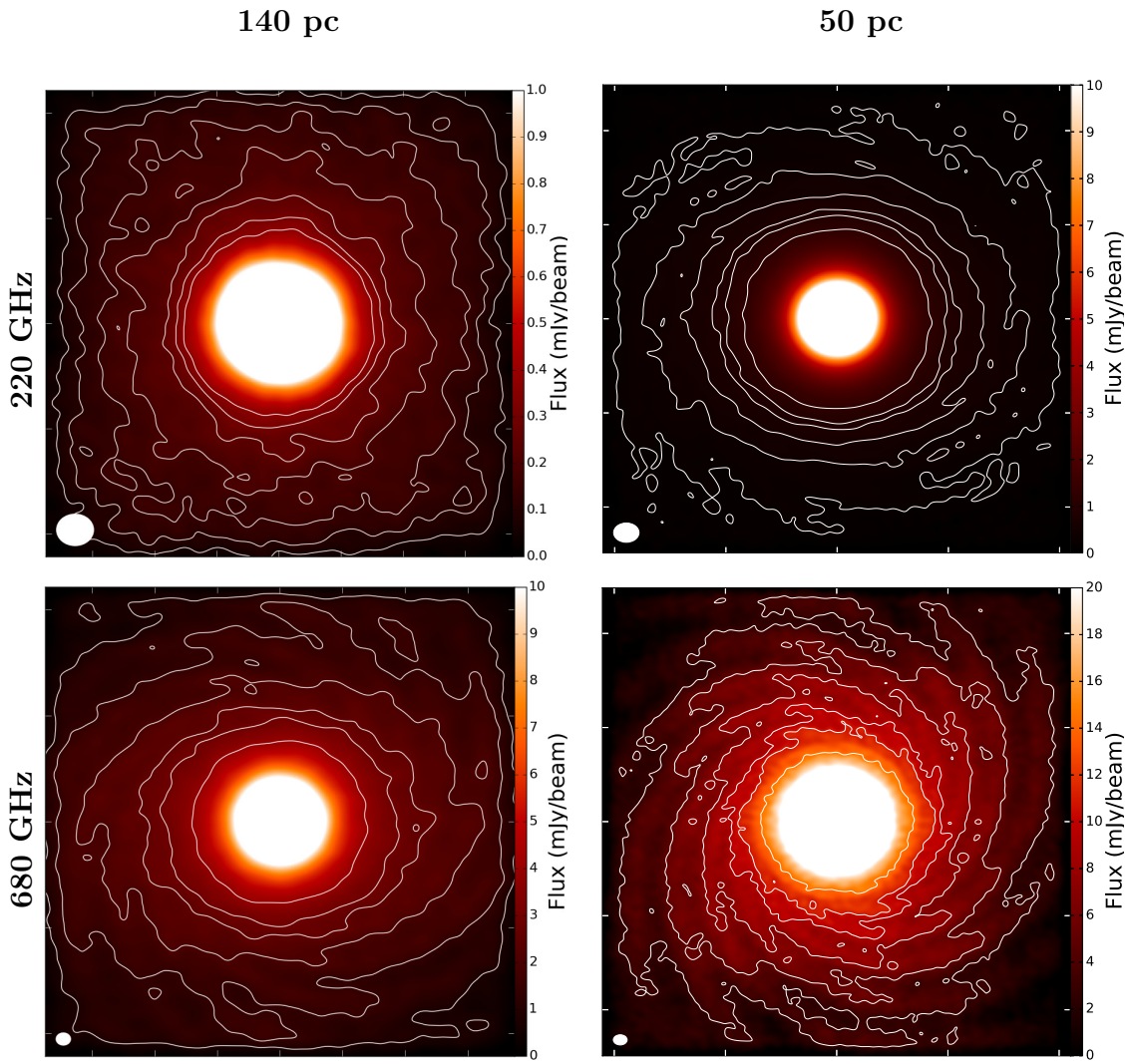


Figure 6.9: All discs are an outer radius of 100 AU, imaged at 220 GHz and 680 GHz, at a distance of 50pc and 140 pc. $M_* = 1M_\odot$, $R_* = 2.325\ R_\odot$ and $T_* = 4350$ K. Right column shows the disc imaged at a distance of 50 pc, as if it were TW Hydrae, in the Hydra region. Left column shows discs imaged at a distance of 140 pc, as in the Taurus region. Contours are at intervals of the RMS, given in Table 6.1, beam sizes are also in Table 6.1.

both the cooling time and α to vary locally, so the relative strengths of our perturbations are much less in the outer part of the disc than they would be in an SPH simulation where β is fixed at some relatively low value.

6.5.4 Comparison With Observed Systems

Currently, there aren't any observations that are directly comparable to what we present here. There are, however, some systems with spiral features typically observed in NIR scattered light. We consider the properties of such systems and compare them to what we suggest would be required if self-gravity is to be the source of the spiral features.

6.5.4.1 MWC 758

The transition disc around Herbig A5 star MWC 758 is located in the edge of the Taurus star forming region at a distance of 279^{+94}_{-48} pc (van Leeuwen, 2007). It is 3.5 ± 2 Myr old (Meeus et al., 2012)and has a stellar mass of 2 $M_\odot$ (Isella et al., 2010). The disc mass and radial extent are approximated from sub-millimetre observations as 10^{-2} $M_\odot$ and $\sim$ 100 AU respectively (Andrews et al., 2011a). The accretion rate is estimated as somewhere between 2×10^{-7} $M_\odot$ yr^{-1} (Isella et al., 2008) and 1×10^{-8} $M_\odot$ yr^{-1} (Andrews et al., 2011a).

The first near-IR (NIR) scattered light images, clearly showing the discovery of spiral arms, were given in Grady et al. (2013), obtained using Subaru and 1.1 μm Hubble Space Telescope/NICMOS data. The parameterised fit of the spiral arms was performed by Grady et al. (2013) following Muto et al. (2012). It is possible that such spirals are launched by a perturbing body, if so it would require a mass

of ~ 5 M_J, which is consistent with continued accretion onto the central star.

Marino et al. (2015) combine VLA Ka and ALMA maps to show that the disc is clearly non-axisymmetric. The disc is fit with a steady state vortex solution to explain the spiral arms. The authors suggest that the compact emission in VLA Ka data is consistent with an accreting companion object such as a forming planet, which could also be responsible for the spiral arms imaged in scattered light. The companion planet scenario is consistent with simulations conducted by Dong et al. (2015a).

Similarly, MWC 758 was imaged in scattered light by Benisty et al. (2015b), using VLT/SPHERE to achieve a higher resolution than previously achieved. The spirals arms were again modelled using density wave theory, with two planetary companions launching the spiral arms. Although the spirals are interpreted as being due to planetary companions, other mechanisms, such as GI, can launch spiral waves with low m modes that are capable of matching these observed features, as shown by Dong et al. (2015b). The measured disc mass (10^{-2} $M_\odot$) of MWC 758 is probably too low to trigger gravitational instabilities (see e.g. Gammie 2001), however, as discussed in our introduction, there are large uncertainties in the ratio of dust to gas and there is evidence that T-Tauri disc masses have been systematically under-estimated.

Observations have revealed a complex morphology of the disk of MWC 758. To understand the origin of these spiral features, both modelling and high resolution images in the sub-mm with ALMA is needed. Scattered light traces the surface variations in a disc (a valid assumption for vertical isothermal hydrostatic

equilibrium), whilst to probe structures near the midplane it is preferable to use longer wavelengths with high spatial resolution.

In this work, we model MWC 758 as if it is self-gravity that is responsible for these spirals, to see if it is indeed the likely origin of these features. We simply assess whether it exists in the parameter space required for self-gravity to exist.

We enter into our model a host star with mass 2 $M_\odot$, a disc outer radius of 100 AU and an accretion rate of 2×10^{-7} $M_\odot$ yr^{-1}. In order for the disc to be in a quasi-steady, self-gravitating state for these specific parameters requires a disc-to-star mass ratio of $q \sim 0.25$. This gives a total disc mass that is over an order of magnitude larger than that given by Andrews et al. (2011a). This means that either:

1. The spirals are due to self-gravity, and the mass of the disc surrounding MWC 758 has been underestimated by a factor of 50. Even if this is the case, Dong et al. (2015b) have recently shown that self-gravitating spiral arms obey $m \sim 1/q$, suggesting that the expected dominant m-mode would be 4, if the spirals are due to self-gravity. However, for $m = 2$ spiral modes to dominate typically requires $q \gtrsim 0.5$, and that the accretion rate be high, of order $\sim \dot{M} \approx 10^{-6}$ $M_\odot$ yr^{-1} (Dong et al., 2015b). Such a disc would have non-local angular momentum transport (Forgan et al., 2011), and as such would not be well-described by the viscous approximation in our analytical model.

2. The disc is self-gravitating, but the accretion rate is much lower than any of the measured values given by Andrews et al. (2011a) for MWC 758, and

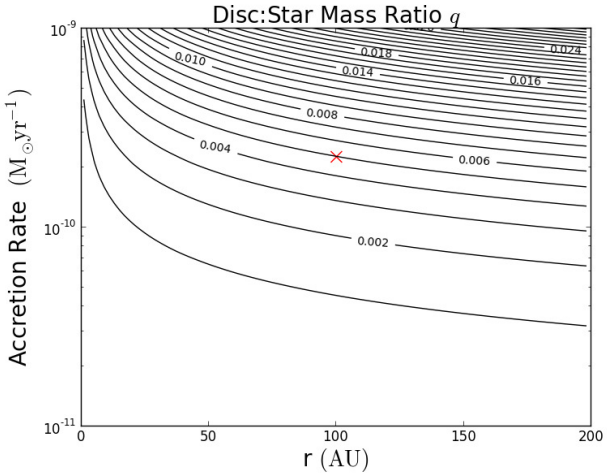


Figure 6.10: Contour plot of disc-to-star mass ratio for accretion rate and radius, plotted in the range required for a host star of 2 $M_\odot$ to have a quasi-steady, self gravitating disc of mass ratio $q = 0.005$, or disc mass $M_D = 0.01$ $M_\odot$. The position of MWC 758 is marked by a red cross. In order for the system to be in a self-gravitating, quasi-steady state with a measured disc mass of $M_D = 0.01$ $M_\odot$ requires an accretion rate of $\sim 10^{-10}$ $M_\odot$ yr^{-1}.

the measured disc mass is correct. Figure 6.10 shows that for a disc around a host star of 2 $M_\odot$ to have a mass of 10^{-2} $M_\odot$ (or equivalently $q \sim 0.005$) requires that the accretion rate be of order $\sim 10^{-10}$ $M_\odot$ yr^{-1}. If this is the case, it is highly unlikely that spiral structure would be detectable since the $\alpha_{\rm grav}$, and therefore perturbation strength of the spiral, would be incredibly low.

3. Both the disc mass of MWC758 and the accretion rate are accurate. The spiral structure visible is due to some other mechanism, perhaps planet - disc interaction as discussed in Benisty et al. (2015b), and not due to self-gravity. Accretion is therefore driven by something other than GI, such as MRI.

Gravitational instabilities are certainly capable of producing structures which match the morphologies of observed low m-mode systems. However, it also makes demands on the system that in the case do not appear to be met, i.e. that disc-to-star mass ratio and accretion rate are very high.

Something else to bear in mind is that our analytic models make assumptions that are likely no longer valid in high mass ($q \gtrsim 0.5$) discs, in which global ($m \sim 2$) spiral modes dominate. When global torques are induced, the angular momentum transport is no longer local, and the semi-analytic models using a local viscous approximation are no longer justified.

Additionally, this semi-analytic model uses the midplane cooling time to determine the effective gravitational α. For massive discs, this will be largely underestimated compared to the actual α value in a global, radiatively cooling disc (Forgan et al., 2011). Given that the spirals in MWC 758 appear global in nature might suggest that we can't use our semi-analytic model in this comparison. However, producing such global spirals via GI would require disc properties even more discrepant than our model suggests, and so the basic conclusion would remain unchanged.

This should serve as a word of caution to the analysis of future observations of discs with non-axisymmetric structure. Modelling non-local discs in the local approximation will return discs with spiral amplitudes far lower than would realistically be present. On the other hand, such discs would be extremely massive and have high accretion rates. Not only is it unlikely that they would be confused for lower mass discs, they will also remain in this phase for a very short time. The

local approximation is therefore probably reasonable for anything that is likely to be observed by ALMA.

Adding irradiation to MWC 758 would allow the system to maintain a larger disc mass. However, as shown in Section 6.5, this removes spiral features from the disc. Adding just enough irradiation to the disc so that spiral features are still observable does not change the result we get for a system with parameters matching those of MWC 758. If the spirals present in MWC 758 are due to disc self-gravity, then they cannot be modelled using a local prescription of angular momentum transport, but this would also require the system to have very different properties to those observed.

6.5.4.2 SAO 206462

SAO 206462 is an isolated 1.7 $M_\odot$ Herbig Ae/Be star at a distance of 142 pc in the constellation Lupus (Müller et al., 2011). It has a $\sim 10^{-3}$ $M_\odot$ disc (Thi et al., 2001) and an accretion rate of 5.37×10^{-9} $M_\odot$ yr^{-1} (Garcia Lopez et al., 2006) and an outer radius of 140 AU. Scattered light observations in NIR have revealed spiral structure in the outer disc (Garufi et al., 2013; Muto et al., 2012), and sub-mm ALMA observations have revealed large-scale asymmetries in the dust continuum (Pérez et al., 2014). These asymmetries have been fit using a vortex prescription, following Regly et al. (2012) by Pérez et al. (2014), however, those authors concluded that the vortex prescription did not reproduce every observed feature, and significant residuals remained which coincided with the spiral arms seen in H-band scattered light.

Although the disc mass is probably too low to trigger gravitational instabilities,

as is the case with many T-Tauri stars there is evidence for systematic underestimation of the disc mass. We model SAO 206462 as if disc self-gravity is responsible for the spiral features present in the disc. Using our model, to match the accretion rate of SAO 206462 requires a disc-to-star mass ratio of $q \sim 0.1$ in order for the disc to be in a quasi-steady, self gravitating state.

Spiral arms which have GI as their origin make additional demands on a system that in this case do not seem to be fulfilled. The spirals must be compact (on scales less than ~ 100 AU), the disc must be massive ($q \gtrsim 0.25$) and the accretion rate must be high (Dong et al., 2015b). This leaves us with the following available conclusions:

1. The disc mass has been underestimated by several orders of magnitude, and the disc for SAO 206462 is actually well within the self-gravitating regime. However, even if this is the case, such a disc would not produce clear spiral structure due to the low $\alpha_{\rm grav}$ and therefore spiral amplitude.

2. The accretion rate is much lower than that measured for SAO 206462, and the disc mass measured is correct. This would further decrease the amount of flux from the disc and again cause difficulty observing it.

3. Both the disc mass and accretion rate are accurate, and the spiral features are not due to disc self-gravity.

6.5.4.3 HD 142527

The transition disc HD142527 has been observed in the near-IR, and has been revealed to have a unique morphology, appearing to consist of two bright arcs

facing each other and one spiral arm (Fukagawa et al., 2006). The central star's mass and age are respectively estimated at $1.9-2.2\mathrm{M}_{\odot}$ and $1-12$Myr (Fukagawa et al., 2006; Verhoeff et al., 2011). It has an accretion rate of $6.92\times10^{-8}\mathrm{M}_{\odot}$ yr^{-1} (Garcia Lopez et al., 2006) and the estimated flow rate of gas in the gap in the disc is between 7×10^{-9} and $2\times10^{-7}\mathrm{M}_{\odot}$ yr^{-1} (Casassus et al., 2013). The total disc mass has been measured from gas-to-dust ratios as ~ 0.1 M$_{\odot}$ (Verhoeff et al., 2011).

Spiral arms have been imaged in ^{12}CO $J=2-1$ and $J=3-2$ using ALMA by Christiaens et al. (2014), who placed lower limits on the mass of each spiral arm at $\sim 10^{-6}\mathrm{M}_{\odot}$. These features were interpreted as an acoustic wave launched by a planet (see e.g. Muto et al. 2012), however since it is now thought probable that HD 142527 has a low-mass stellar companion (Biller et al., 2012), the spiral structures could certainly be tidally induced. GI is an alternative scenario which is able to replicate this grand design spiral structure, and since both Christiaens et al. (2014) and Fukagawa et al. (2013) find $Q\sim2.0$, there is evidence for gravitational instability being responsible for the spiral structure present.

We assume a 2 M$_{\odot}$ central star, and an accretion rate of $6.92\times10^{-8}\mathrm{M}_{\odot}$ yr^{-1}. Since there is evidence that the disc may extend as far out as 600 AU in radius (Christiaens et al., 2014), we extend our disc out to 600 AU.

To match the observed accretion rate of HD 142527 and extend out to 600 AU, such a disc would require an incredibly high disc-to-star mass ratio of $q\sim0.75$. Such a disc would certainly have incredibly high global torques, and in reality would probably not survive in this quasi-steady, self-gravitating state. Therefore,

the most likely explanation for the spiral structure observed in HD 142527 is not self-gravity.

An alternative explanation for the spiral structure is tidal interaction due to its low mass companion ($M_{\rm companion} \sim 0.1$ M$_\odot$), which is potentially on an eccentric orbit around HD 142527 (Baines et al., 2006; Biller et al., 2012; Close et al., 2014; Fukagawa et al., 2006).

6.5.4.4 Conclusions from Observed Systems

We examined the parameter space of three transition discs to determine if the non-axisymmetric structure which has been imaged in those discs could feasibly be due to disc self-gravity. For all three systems, it seems unlikely, unless the disc mass has been significantly underestimated. Even if the disc mass has been underestimated, and the disc is self-gravitating, we may expect to see a different number of m-modes dominant in the disc. Self-gravity imposes additional requirements on a system which do not seem to be consistent with the parameters of these systems.

6.6 Connecting SPH Simulations to Observations

6.6.1 Author's Note

The work that appears in section 6.6 was published in Dong et al. (2015b). CH was the *second* author in this work, responsible for the analysis of the physical properties of the SPH simulations. Ruobing Dong carried out the MCRT calculations, and produced the synthetic images.

6.6.2 SPH and MCRT Simulation Setup

Up to now, we have established that our analytical model poorly describes systems that are dominant in the low m-modes. Additionally, we have generated synthetic observations at longer wavelengths than the systems we are comparing to are imaged at. With this in mind, we run 3 global SPH simulations to calculate the gas density structures of gravitationally unstable discs, and generate synthetic images using the Whitney MCRT code.

The code includes the hybrid radiative cooling method of Forgan et al. (2009). The details of the algorithm are given in Forgan et al. (2009), but the general idea is to merge the polytropic cooling method of Stamatellos et al. (2007) with the flux-limited diffusion of Mayer et al. (2007b). The polytropic cooling handles the energy loss in the system, and flux-limited diffusion handles the positive energy exchange between particles.

With this in mind, we have combined 3D SPH simulations with MCRT calculations, conducted using the Whitney MCRT code. A disc was constructed in 3D spherical coordinates, with 400 radial cells binned logarithmically, 200 polar cells and 256 azimuthal cells. Since we do not include stellar irradiation in our simulations, the outer regions of the disc can be artificially cold. This was corrected for by readjusting the vertical temperature profile using the temperature calculated from the MCRT code, rather than the SPH temperature. The inconsistency is not serious, it would only cause a slight weakening of the spiral structure. The grain distribution is a power law between 0.02 and 0.25 μm, at these small grain sizes the dust and gas is well-mixed.

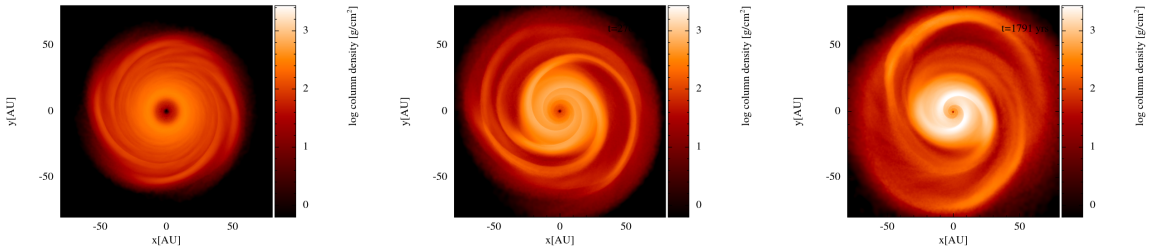


Figure 6.11: Left: MD0125. Centre:MD025. Right: MD05. All discs became gravitationally unstable and developed spiral structure. We can see that grand-design type spirals are present for the largest disc masses.

We produce synthesised polarized intensity images at 1.6 μm for our three discs, and convolved them with a Gaussian point-spread function to achieve an angular resolution of 0.05″ assuming a distance of 140 pc. For reference, the VLT, in single telescope mode, can achieve a resolution of $\sim$ 0.05″.

6.6.3 SPH Results

The final surface density profiles of the disc are shown in Figure 6.11, left is MD0125, centre is MD025 and right is MD05. All discs developed GI, and the left hand panel of Figure 6.12 shows (from top to bottom) the radial surface density profile of the three discs, the midplane SPH temperature, and finally the Toomre parameter in the bottom panel. We find that the temperature in the outer regions of the disc, if taken from the SPH simulations, can be artificially low, and, therefore, the disc is artificially thin. This is because we do not include stellar radiation in our SPH simulations. We therefore correct the vertical density profile of our discs using the MCRT temperature, and the corrected radial profiles, using the MCRT temperature, are shown in the right hand panel of Figure 6.12.

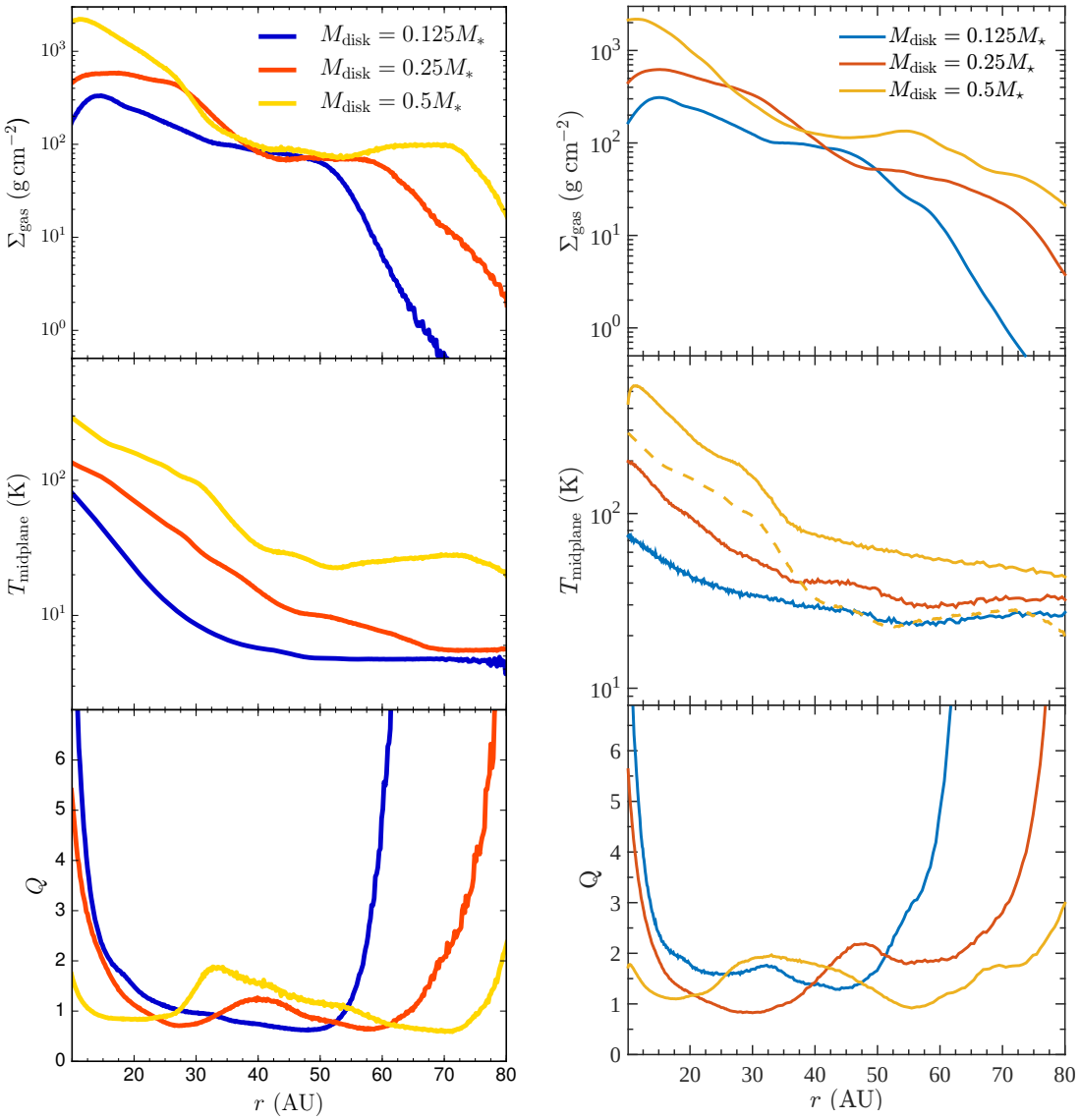


Figure 6.12: Left: Uncorrected radial profiles using SPH temperature. Right: Corrected radial profiles using MCRT temperatures. Both: Top panel shows surface density, middle panel shows midplane temperature, bottom panel shows Toomre parameter. The dashed curve in the right, middle panel is the midplane SPH temperature for the $M_{\rm disc} = 0.5\,M_*$ simulation. Note that the MCRT temperature can be up to a factor of 2 higher than the SPH temperature in the outer disc, this is because we do not use stellar irradiation in the SPH simulation.

The bottom right hand panel in Figure 6.12 shows that as q increases, the gravitationally unstable region (i.e. where $Q < 2$) expands. For MD0125, this is $\sim 20 - 50$ au, for MD025, $\sim 20 - 60$ au, and for MD05 this is $\sim 10 - 70$ au. We find that as the disc settles into a state of quasi-equilibrium, as q increases, the dominant azimuthal wavenumber decreases, with the approximate relationship $m \sim 1/q$. This is consistent with previous work, which shows that more massive discs dominate in the low m-modes (Lodato & Rice, 2004). The model properties are given in Table 6.3, which details the disc-to-star mass ratio, the accretion rate onto the central star, and the dominant azimuthal wavenumber.

Model	q	$\dot{M}$	Dominant Azimuthal Wavenumber
MD0125	0.125	7×10^{-8}	$m \sim 4 - 8$
MD025	0.25	5×10^{-7}	$m \sim 4$
MD05	0.5	5×10^{-6}	$m = 2$

Table 6.3: Final properties for our three non-fragmenting discs. We find that to sustain global, low m-mode spirals requires a high disc mass and high accretion rate. We find an empirical relationship between disc-to-star mass ratio and dominant azimuthal wavenumber such that $m \sim 1/q$.

6.6.4 Connection to Observations

Figure 6.13 shows the surface density image of MD0125, left, and the convolved MCRT image, right. We can see, immediately, our first conclusion: when $q \lesssim 0.1$, spiral arms due to gravitational instability are weak in scattered light, with contrasts that are sufficiently low so as to make them difficult to observe.

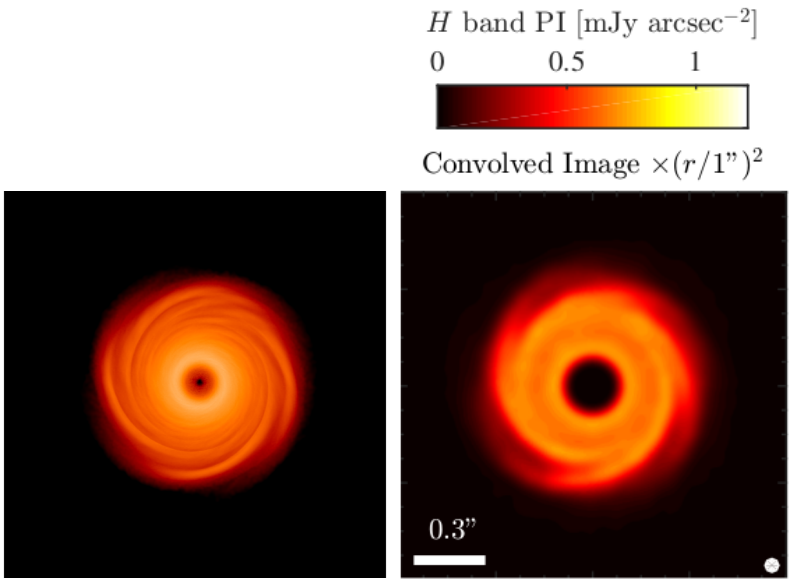


Figure 6.13: Left is the surface density of image of MD0125, right shows H band PI image, face on, scaled by r^2. Distance is assumed to be 140 pc, and the PI image has been convolved with a Gaussian of FWHM = $0.05''$. The inner 14 au ($0.1''$) is masked out.

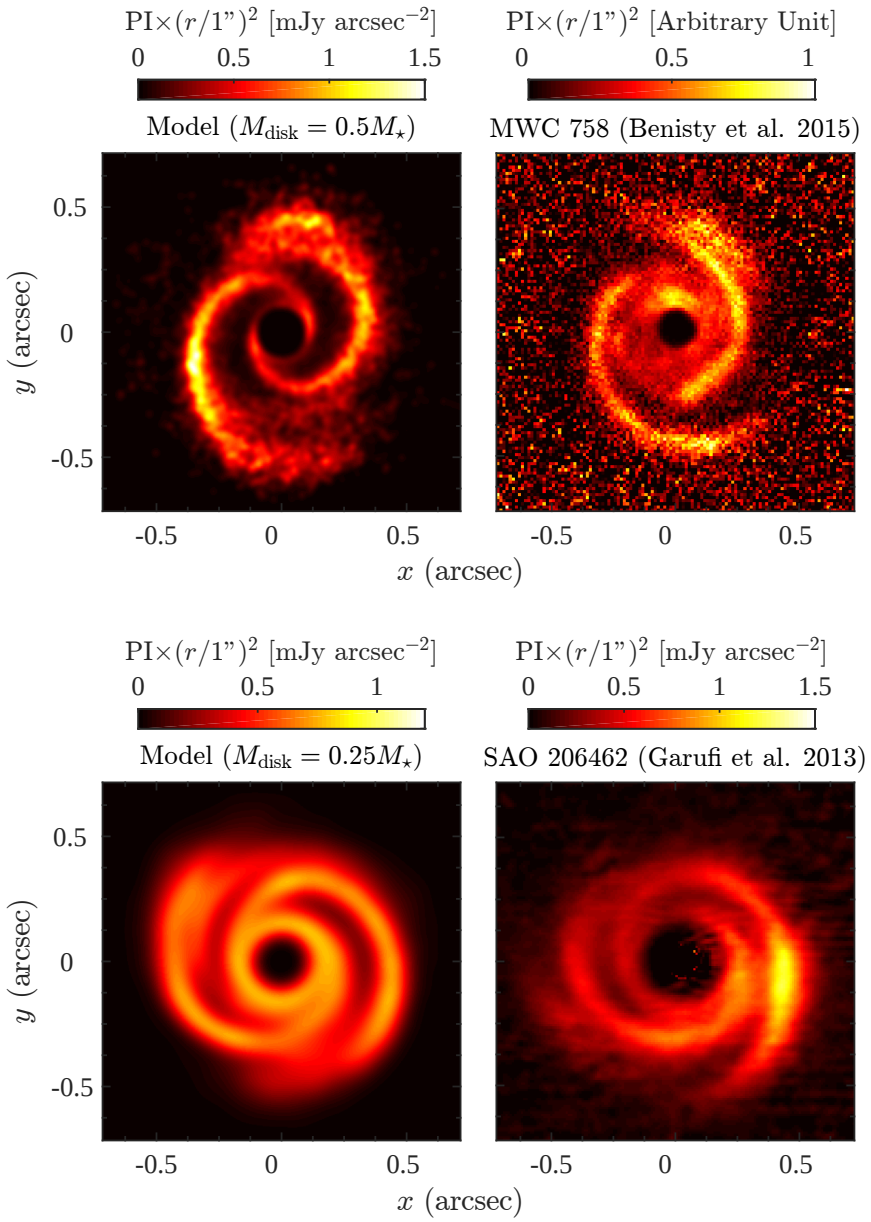


Figure 6.14: Top left panel: synthesised 1.6 μm image of the $q = 0.25$ M_* disc, convolved with a Gaussian PSF of FWHM=0.03″, and assumed to be at a distance of 140 pc. Top right panel: for comparison, the MWC 758 image by Benisty et al. (2015b) (reproduced with permission). Bottom left panel: synthesised 1.6 μm image of the $q = 0.5$ M_* disc, convolved with a Gaussian PSF of FWHM=0.09″, and assumed to be at a distance of 140 pc. Bottom right panel: for comparison, the SAO 206462 image by Garufi et al. (2013) (reproduced with permission). The units in all images are mJy arcsec^{-2} , except for the image of MWC 758, for which the units are arbitrary. We can see the similarity in morphology between our SPH simulations and these images. The inner 14 au is masked out in the synthetic images.

In section 6.5.4, we describe two systems that have been imaged in NIR scattered light, MWC 758 and SAO 206462. We noted that they are not strictly comparable to the continuum emission we modelled, however, they are comparable to the NIR scattered light at 1.6 μm we now consider. We note that the MD05 disc morphologically resembles the MWC 758 system, and the MD025 disc morphologically resembles the SAO 206462 system. This is shown in Figure 6.14, top left shows our scattered light image of MD05, and top right shows the MWC 758 image obtained by Benisty et al. (2015b). Bottom left shows our scattered light image of MD025, bottom right shows the scattered light image of SAO 206462 obtained by Garufi et al. (2013). This leads us to our second conclusion, that GI can produce prominent spiral arms in NIR scattered light, so long as $q \gtrsim 0.25$.

6.7 Discussion and Conclusion

We performed an examination of the parameter space in which self-gravitating discs can exist, using a semi-analytical approach. We generated synthetic observations with the intention of investigating the range of accretion rates, disc masses and disc radii that would allow non-axisymmetric structure to be detected by ALMA. Our intention was not to reproduce the exact morphology of the observations, but rather to understand the strength of a perturbation required to generate an observable spiral arm.

Analytical models using a viscous prescription that assumes local angular momentum transport poorly describe systems which are dominant in the low m spiral modes. Modelling non-local discs in the local approximation returns

spiral amplitudes far lower than would be present in reality. If a quasi-steady, self-gravitating disc can be described analytically using local transport, then there exists a small range of accretion rates for a given radius where the gravitational stress is high enough to generate observable spirals, but not so high as to cause the outer regions of the disc to fragment. However, non-local transport only becomes significant in discs with masses above half the mass of the central star, and such discs will probably have very short lives (Lodato & Rice, 2005; Rice et al., 2010), so our analysis here is probably reasonable for anything that would be detected by ALMA.

Another important factor is external irradiation. If the accretion rate is close to the fragmentation limit, a small amount of external irradiation ($\sim$ 10 K) may prevent fragmentation with increasing accretion rate. If the accretion rate is well below the fragmentation limit, a small amount of irradiation ($\sim$ 10 K) causes the surface density profile of the outer part of the disc to be restructured, as the equilibrium disc structure is now more massive for a given accretion rate and radius. If infall from a natal cloud is occurring, this could well be a trigger for fragmentation, as it is likely that in these regions the Jeans criterion would be satisfied. A moderate amount of irradiation ($\sim$ 30 K) can suppress fragmentation up to higher accretion rates, but at the cost of non-axisymmetric structure.

We performed 3D MCRT calculations with geometries from SPH discs, and found that GI can produce prominent spiral arms in NIR scattered light images, so long as $q \gtrsim 0.25$. The morphology of these discs is strikingly similar to that of observed systems such as MWC 758 and SAO 206462. However, the estimated disc-to-star mass ratio for these systems is $q \sim 0.01$ (Andrews et al., 2011b). If this is correct,

then it is highly unlikely that either of these systems would be susceptible to disc self-gravity. On the other hand, the disc mass estimates for these systems are performed under specific assumptions about dust opacity, and dust-to-gas mass ratio. One, or both, of these assumptions may be incorrect. If so, these systems may have the necessary properties to be self-gravitating.

An additional consideration for GI as an explanation for the spiral arms visible in these systems is the stellar accretion rate. For MD025 and MD05, we measure $\dot{M} = 5 \times 10^{-7}$ M$_\odot$ yr^{-1} and $\dot{M} = 5 \times 10^{-6}$ M$_\odot$ yr^{-1} respectively. However, the measured accretion rate for these systems is of order $\sim 10^{-8}$ M$_\odot$ yr^{-1}(Andrews et al., 2011a). Accretion rates this low would be unlikely to correspond to self-gravitating systems, since, roughly speaking, we require $\dot{M} \gtrsim 10^{-7}$ M$_\odot$ yr^{-1} to have sufficiently large spiral amplitudes so as to be observable. However, accretion in gravitationally unstable discs can be highly episodic. As shown in Figure 3 of Vorobyov & Basu (2005), the accretion rate is highly variable, and for a sufficiently large core mass will frequently drop below $\sim 10^{-8}$ M$_\odot$ yr^{-1}.

Ultimately, our results suggest that there is a relatively small range of parameter space in which a disc could be self-gravitating, not undergo fragmentation, and have spiral amplitudes large enough to be observable by ALMA. Broadly speaking, we would expect the disc mass to exceed 0.1 M$_\odot$, the accretion rate to satisfy $10^{-7} \lesssim \dot{M} \lesssim 10^{-6}$ M$_\odot$ yr^{-1} and the outer radius to be not much more than 100 au. Additionally, the observing frequency and distance to the source also plays a role. We are more likely to observe spiral waves at 680 GHz than at 220 GHz, and it becomes increasingly difficult as the source distance increases.

We also conclude that, in the MWC 758 and SAO 206462 systems, it is unclear whether their spiral features may be attributed to disc self-gravity. It may be the case that another explanation, such as excitation of spiral waves by planets as shown in Dong et al. (2015a), is more plausible.

Although self gravitating discs can certainly match the morphology of observed systems, they also impose strict additional conditions which may not be met. In essence, our analysis suggests that there is a relatively small region of parameter space in which self-gravity may produce observable spiral features. We would therefore caution against interpreting such features as being due to disc self-gravity unless the disc is likely to fall into this region of parameter space.

6.8 Acknowledgements

CH warmly thanks Giovanni Dipierro and Guillaume Laibe for their insightful and elucidating discussion with her concerning dust grains in this model. KR and BB gratefully acknowledge support from STFC grant ST/M001229/1. DF gratefully acknowledges support from the ECOGAL ERC advanced grant. Some calculations for this paper were performed on the University of Exeter Supercomputer, a DiRAC Facility jointly funded by STFC, the Large Facilities Capital fund of BIS, and the University of Exeter, and on the Complexity DiRAC Facility jointly funded by STFC and the Large Facilities Capital Fund of BIS. TJH acknowledges funding from Exeter's STFC Consolidated Grant (ST/J001627/1).

It is a slightly arresting notion that if you were to pick yourself apart with tweezers, one atom at a time, you would produce a mound of fine atomic dust, none of which had ever been alive but all of which had once been you.

Bill Bryson, A Short History of Nearly Everything

7

The Role of Dust Enhancement in Direct Observation of Self-Gravitating Spiral Waves

7.1 Author's Note

In the previous chapter, we showed, using our simple analytical model, that there exists a very small range of accretion rates where the disc stress is large enough to generate observable spirals, but not so large as to trigger disc fragmention.

However, our analysis was, perhaps, overly simplistic, since we assumed that the dust was well-coupled to the gas. In reality, the amount of dust, of varying sizes, depending upon their Stokes number, can collect at pressure maxima in the disc. This will alter the contrast ratio in continuum emission, between the spiral arms and inter-arm regions.

In this chapter, we modify the analytical model, discussed in Chapter 6, to include a "dust-trapping" effect, whereby the amount of dust, of a given grain size, is enhanced in the spiral arm, based upon the Stokes number of the grain. We show that with the inclusion of this important behaviour in our analytical model, we are able to detect self-gravitating spiral waves over a wider range of accretion rates.

From this, we conclude that in order to detect self-gravitating spiral waves in protostellar discs, we probably require a significant amount of grain growth to have taken place in the disc.

7.2 Introduction

There are two complementary theories of planet formation, the first, that we have already discussed in this thesis, is that of direct formation through gravitational instability. This is limited to the outer regions of discs, typically beyond ~ 50 au, where the disc may be susceptible to gravitational fragmentation. Gravitational instability may offer a formation mechanism to wide-orbit planets, such as those in the HR 8799 system, since the core accretion theory typically requires timescales that are longer than the lifetime of the disc to form such planets.

The second, and most widely accepted, theory of planet formation is that of core accretion (CA). In this scenario, micron sized dust grains coagulate into larger particles, before settling into the disc midplane, where they accrete enough material to grow into kilometre-sized planetesimals. The planetesimals grow via collisions into planetary cores, and if they are sufficiently massive, may subsequently accrete a gaseous envelope, ultimately forming a gas-giant planet.

The first stage of planet formation in CA, the formation of planetesimals through the coagulation of particles, is beset by several problems, the most famous of which is the *metre barrier*, where particles ~ 1 m in size have the largest radial drift velocity (Weidenschilling, 1977), and so particles are likely to be destroyed before they reach much more than ~ 1 m in size. Although *pebble accretion theory* (Lambrechts & Johansen, 2012; Levison et al., 2015) looks to be a promising solution, the theory may still require formation timescales longer than typical disc lifetimes (Bodenheimer et al., 2000; Pollack et al., 1996). Radial drift, for all particles, is itself a problem. From the momentum equation for an unmagnetised, inviscid fluid, we have

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\frac{1}{\rho}\nabla P - \nabla \mathbf{\Phi}. \tag{7.1}$$

In cylindrical coordinates, we then have, for the orbital velocity of the gas,

$$\frac{v^2_{\phi,\mathrm{gas}}}{r} = \frac{\mathrm{G}M_*}{r^2} + \frac{1}{\rho}\frac{dP}{dr} \tag{7.2}$$

Since pressure decreases with radius in Keplerian discs, the pressure gradient is negative, meaning that the gas typically orbits at sub-Keplerian velocities.

Dust particles, being solid bodies, do not feel this pressure force and so orbit at Keplerian velocity. Hence, they feel a headwind, lose angular momentum via drag, and begin to drift inwards.

This inward radial drift presents a problem for planet formation, since too high an inward velocity would lead to dust grains being destroyed by radiation from the central star before they have undergone significant growth. However, a potential solution to this may come from the complementary theory of planet formation, gravitational instability. If we suppose that the disc is self-gravitating, and does not cool rapidly enough to fragment, but instead reaches a state of quasi-steady equilibrium, then we can expect the disc to sustain long-lived spiral arms.

In the presence of these spiral arms, if we begin at an inter-arm region, and head radially outwards, we will see an increase in pressure until we reach the most dense point of the spiral arm, and then a decrease in pressure as we pass it and continue outwards. This means that, locally, the pressure gradient varies from positive to negative, meaning that the gas orbits at super-Keplerian and sub-Keplerian velocities respectively, with the dust experiencing a "push" force from the super-Keplerian gas, and a pull force from the sub-Keplerian gas. We may then expect dust particles to become collected at local pressure maxima. This effect has been shown in 3D global hydrodynamics simulations by Rice et al. (2004), and may, potentially, offer a mechanism for accelerated planetesimal growth in protostellar discs.

In this chapter, we develop an analytical prescription for varying the density of grains of given sizes based upon their Stokes number (defined and discussed

in section) at a particular location in the disc. It is simple, and allows for grain growth such that the grain distribution may not accurately be described by the usual power-law distribution. However, our purpose here is to show the importance of dust-enhancement, if we are to detect spiral features that are due to disc self-gravity, rather than accurately represent the dust distribution typical in protostellar discs.

7.3 Model

The geometry of the model is described in chapter 6, section 6.2 and we refer the reader to this chapter for an in depth explanation. However, the basic idea is that for a self-gravitating disc in a quasi-steady state, that can be described using the viscous, local approximation, then for a given radius, and accretion rate, there is just one value of the surface density that can describe this system. This is determined through successive iteration until the measured accretion rate in the system, $\dot{M}$, is equal to the imposed accretion rate (to a small tolerance), both of which are described by

$$\dot{M} = \frac{3\pi\alpha c_{\rm s}^2 \Sigma}{\Omega}, \tag{7.3}$$

where α is the effective gravitational stress, $c_{\rm s}$ is the sound speed, Σ is the disc surface density and Ω is the orbital frequency. Cossins et al. (2009) find an empirical relationship for the amplitude of the spiral density waves, such that

$$\frac{\langle \delta\Sigma \rangle}{\langle \Sigma \rangle} \approx \beta^{-\frac{1}{2}}, \tag{7.4}$$

and we sinusoidally impose this around the disc, so

$$\delta\Sigma(\phi) = -\langle\delta\Sigma\rangle \cos(m\phi), \tag{7.5}$$

where m is the azimuthal wavenumber selected by the user. Considering that high mass discs are dominated by low m-mode spirals (Lodato & Rice, 2005), and the recent empirical results by Dong et al. (2015b), who find that $m \sim 1/q$, we apply this and vary m as the disc mass changes with accretion rate. However, we note that strictly speaking, any low m-mode structure will likely exert considerable global torques in the disc. If this is the case, then our local, viscous approximation may underestimate the density amplitude of these spiral arms.

7.3.1 Dust Trapping

There is no clear-cut method for developing an analytic prescription for dust surface density that removes time dependence. As previously mentioned, our aim in this chapter is not to exactly reproduce the dust surface density profile of a protostellar disc, but, rather, to develop an understanding as to where we may expect dust trapping to occur in our model. With this in mind, we begin by defining the Stokes number as a dimensionless stopping time, such that

$$St = \frac{t_{\rm s}}{t_{\rm obstacle}}, \tag{7.6}$$

where $t_{\rm s}$ is the stopping time of the dust particle, and $t_{\rm obstacle}$ is the characteristic time of the obstacle, which in this case is the gas in the disc. As the dust particle traverses through the gaseous medium, it loses momentum due to friction with

the gas. This coupling drag force can be described through (Weidenschilling, 1977; Whipple, 1972)

$$F_{\rm D} = -\frac{1}{2} C_{\rm D} \pi a^2 \rho_{\rm g} u^2, \tag{7.7}$$

where $C_{\rm D}$ is the drag coefficient, πa^2 is the cross section of the particle, $\rho_{\rm g}$ is the density of the gas and u is the relative velocity of the particle. The coefficient $C_{\rm D}$ depends upon the drag regime, and is given by

$$C_{\rm D} = \begin{cases} \frac{8}{3}\frac{c_{\rm s}}{u} & a < \frac{9\lambda}{4} \text{(Epstein regime)} \\ \frac{24}{Re} & Re < 1 \\ \frac{24}{Re^{0.6}} & 1 < Re < 800 \\ 0.44 & Re > 800, \end{cases} \tag{7.8}$$

where Re is the Reynolds number, and λ is the mean free path of the gas particles, given by

$$\lambda = \frac{1}{n\sigma} = \frac{m_{\rm H_2}}{\rho_g \pi a_{\rm g}^2}, \tag{7.9}$$

where n is the number density, σ is the cross-sectional area, given by $\pi a_{\rm g}^2$, of the gas particles, $\rm H_2$ is the mass of each gas particle assuming it is composed of molecular hydrogen. The dimensionless Reynold's number, which characterises the turbulence of the flow, is given by

$$Re = \frac{2a\rho u}{\eta}, \tag{7.10}$$

where η is the gas viscosity, given by[18]

$$\eta = \rho_{\rm g}\nu. \tag{7.11}$$

So long as this viscosity is collisional, then $\nu = \frac{\rho c_{\rm s}\lambda}{2}$, and the Reynold's number can then be expressed as

$$Re = \frac{4au}{\lambda c_{\rm s}}. \tag{7.12}$$

The stopping time can then be calculated by

$$t_s = \frac{m_{\rm p}u}{F_{\rm D}}, \tag{7.13}$$

which, in the Epstein regime, gives

$$t_s = \frac{\rho_{\rm s}a}{\rho_g c_{\rm s}}, \tag{7.14}$$

Which now puts us in a position to calculate the Stokes number,

$$St = \frac{\rho_{\rm s}a\Omega}{\rho_g c_{\rm s}} \tag{7.15}$$

Realistically, in protostellar discs, the Reynold's number is large enough that we may expect significant turbulence (values between 1-800 are typical in the protostellar disc scenarios that we consider here Rice et al. 2004, the Reynold's

[18]By definition, the kinematic viscosity, ν, is the ratio of dynamic viscosity, η, to the density of the material, such that $\nu = \frac{\eta}{\rho}$. More generally, if we consider two large plates, parallel to each other and separated by a layer of fluid, with one fixed, and one moving parallel at speed u, then the force required to keep the moving plate travelling at constant speed is described by $F = \eta A\frac{\partial u}{\partial y}$, where $\frac{\partial u}{\partial y}$ is the rate of shear deformation.

number is only likely to exceed ~ 800 in the inner disc). If the drag coefficient was calculated by using the Reynold's number, our calculations of the Stokes number would leave us with the problem of a time dependent variable, the particle velocity, in our value of St. However, the mean free path in a protostellar disc is, mercifully, large enough that we may reasonably expect to be in the Epstein regime everywhere in the disc, so long as the size of the particles, a, is less than the mean free path length, λ. Formally, this condition is given by (see, e.g., Weidenschilling 1977; Whipple 1972)

$$a \lesssim \frac{9}{4}\lambda. \tag{7.16}$$

In a protostellar disc, typical values yield $\lambda = \frac{m_{\mathrm{H_2}}}{\rho_g \pi a_{\mathrm{g}}^2} \sim \frac{10^{-24}}{10^{-11}\cdot(10^{-8})^2} = 1000$ cm, which means that so long as our $a \lesssim 2250$ cm, then the drag force on the grains is in the Epstein regime.

The radial velocity of dust, u_{r}, in a protostellar disc has two components, the first is bulk movement due to accretion of the gas onto the central star, and some radial drift velocity with respect to the gas, given by

$$u_{\mathrm{r}} = \frac{u_{\mathrm{r,gas}}}{1+\mathrm{St}^2} - \frac{2u_{\mathrm{n}}}{\mathrm{St}+\frac{1}{\mathrm{St}}} \tag{7.17}$$

where the first term is the drag component, originating from the radial movement of the surrounding gas, moving with velocity $u_{\mathrm{r,gas}}$, and the second term is the radial drift velocity of the particle, which peaks at $St = 1$, with u_{n} denoting the

maximum radial drift velocity of the particle, given by

$$u_{\mathrm{n}} = \frac{1}{2\rho_{\mathrm{g}}\Omega_{\mathrm{K}}}\frac{\partial P_{\mathrm{g}}}{\partial r}, \tag{7.18}$$

where ρ_{g} is the density of the gas, Ω_{K} is the Keplerian angular velocity and $^{\partial P_{\mathrm{g}}}/_{\partial r}$ is the pressure gradient (Weidenschilling, 1977).

The second term in equation 7.17 forms the basis of our analytical prescription for dust trapping. Our reasoning is simple; since the largest radial drift can be found when $St = 1$, it makes sense that the most trapping will occur for particles with $St = 1$. It makes physical sense, then, to guess that the form of the trapping may also follow the same function as the radial drift component of equation 7.17, such that the amount of trapping is proportional to $\frac{2}{St+St^{-1}}$.

To model this effect, we assume the canonical value for the total dust-to-gas mass ratio in the disc, $\frac{M_{\mathrm{dust}}}{M_{\mathrm{gas}}} = 0.01$. This remains constant. We then make use of the empirical results of Dipierro et al. (2015), who, when modelling the evolution of dust surface density using SPH, found that the peak total dust-to-gas ratio in spiral arms is given by

$$\left(\frac{\Sigma_{\mathrm{d}}}{\Sigma_{\mathrm{g}}}\right)_{\mathrm{max}} = 0.02, \tag{7.19}$$

where Σ_{d} is dust surface density and Σ_{g} is gas surface density. Since we expect the dust-to-gas fraction to be highest at pressure maxima, we impose that at spiral arm density peaks, the local dust-to-gas mass ratio is given by

$$\frac{\Sigma_{\mathrm{d}}}{\Sigma_{\mathrm{g}}} = 0.02, \tag{7.20}$$

and at spiral arm density troughs

$$\frac{\Sigma_{\rm d}}{\Sigma_{\rm g}} = 0.001, \tag{7.21}$$

since it is unlikely that the dust would be entirely removed in these regions. It is important to note that we have not changed the overall amount of dust in the disc, merely redistributed it.

At this point, we now allow grains of different sizes to have different "grain fractions", $\mathcal{G}$, at different locations in the disc, such that

$$\mathcal{G}^i = \frac{\Sigma_{\rm d}^i}{\Sigma_{\rm g}}, \tag{7.22}$$

where i is an index that distinguishes these grains from the total dust parameter. We begin by imposing that the *total* dust-to-gas fraction, sinusoidally, such that it peaks at 0.02 in the center of spiral arms, and troughs at 0.001 in the inter-arm regions. We consider $n_{\rm dust}$ different grain sizes, logarithmically spaced in size, between 0.1 μm and 2000 μ in section 7.4.1, and between 0.1 μm and 100 cm in section 7.5. The bins are linearly spaced. We set $n_{\rm dust} = 20$, and do not consider more grains due to the computational cost of solving the radiative transfer equation for so many grain types.

Initially, each grain is given a grain fraction of

$$\mathcal{G}^i = \frac{\Sigma_{\rm d}}{\Sigma_{\rm g}} \cdot \frac{1}{n_{\rm dust}}, \tag{7.23}$$

where we have divided the *local* dust-to-gas ratio, which varies sinusoidally

between arm and inter-arm regions, by the total number of grain types. We then calculate St, using equation 7.15, for each grain at this location in the disc. The initial grain fraction, $\mathcal{G}^i$, is then modified, such that it peaks when the radial drift is at its maximum, so we now have, for the final grain fraction $\mathcal{F}^i$,

$$\mathcal{F}^i = \mathcal{G}^i \cdot \frac{2}{St + \frac{1}{St}}. \tag{7.24}$$

This is then normalised, such that the sum of all the modified grain fractions is equal to the *total local* grain fraction,

$$\sum_i^{n_{\rm dust}} \mathcal{F}^i = \frac{\Sigma_{\rm d}}{\Sigma_{\rm g}}. \tag{7.25}$$

7.3.2 Generation of Synthetic Images

We input our model into the TORUS radiation transport code (Harries et al., 2004; Haworth et al., 2015; Kurosawa et al., 2004) to determine radiative equilibrium, using the Lucy (1999) Monte Carlo (MC) photon packet algorithm. We use a three dimensional adaptive mesh that stores density, temperature and dust properties. In regions of large opacity gradients, the cell is subdivided into children, such that the gradient is well resolved.

For our radiative transfer results, we assume typical values for a pre-main-sequence star, with central source mass of $M_* = {\rm M}_\odot$, $R_* = 2.325\ R_\odot$ and $T_{\rm eff} = 4350$ K. The dust in our model consists of Draine & Lee (1984) silicates,

with a grain size distribution of

$$n(a) \propto a^{-q} \qquad \text{for} \qquad a_{\rm min} < a < a_{\rm max}, \tag{7.26}$$

where $a_{\rm min}$ and $a_{\rm max}$ are the minimum and maximum size of the dust grains. We run all calculations with $a_{\rm min} = 0.1$ μm, and select $a_{\rm max} = 2000$ μm for results in section 7.4.1, so that we may compare our modified model with the one described in chapter 6, and $a_{\rm max} = 100$ cm for results in section 7.5. The power-law exponent is $q_{\rm ism} = 3.5$ (Mathis et al., 1977), and the *total* dust density is 1% of the gas density.

Once we have performed our radiative equilibrium calculations, the emission maps are used as inputs to the ALMA simulator built into CASA (version 4.5.2). We scale disc sizes and flux to a distance of 140 pc, and all images shown in the chapter were generated using the "C40-8.cfg" antennalist. Noise from atmospheric transmission was included, and since we image at 220 GHz only in this chapter, precipitable water vapour (PWV) is set to 1.262 mm.

7.4 Results

Our results section is broken into two parts: first, in section 7.4.1, we consider a maximum grain size of 2000 μm so that we can directly compare our results to the model in Chapter 6, where we do not include any modification to the local grain fraction (i.e. $\Sigma_{\rm d}/\Sigma_{\rm g}$=0.01 everywhere). In section 7.5, we change $a_{\rm max}$ to 100 cm, and examine, for a variety of grain sizes, their abundance in the disc.

7.4.1 Comparison to Previous Results

We begin with four simulated ALMA images, shown in Figure 7.1. All four discs have eight spiral arms, the discs in the top row have an accretion rate of $\dot{M} = 2.8 \times 10^{-7}$ $M_\odot$ yr^{-1}, and in the bottom row the accretion rate is $\dot{M} = 5.0 \times 10^{-8}$ $M_\odot$ yr^{-1}. The discs in the left column have a constant dust-to-gas mass ratio, $\Sigma_d/\Sigma_g = 0.01$ everywhere in the disc, so as to directly compare to results in Chapter 6.

In the right hand column, the grain fraction for each grain size varies as described in section 7.3. All of the discs are images at a distance of 140pc, centered at 220 GHz, with 1 hour of integration time. The beam size is identical for all images, and shown in the bottom left of each image. Beam size is $0.053'' \times 0.048''$. Contours are at intervals of 3.0, 6.0, 9.0, 12.0, 15.0 and 18.0 $\times$ the RMS noise in each image. RMS for each image is: top left 18 μJy beam^{-1}, top right 131 μJy beam^{-1}, bottom left 21 μJy beam^{-1}, bottom right 100 μJy beam^{-1}.

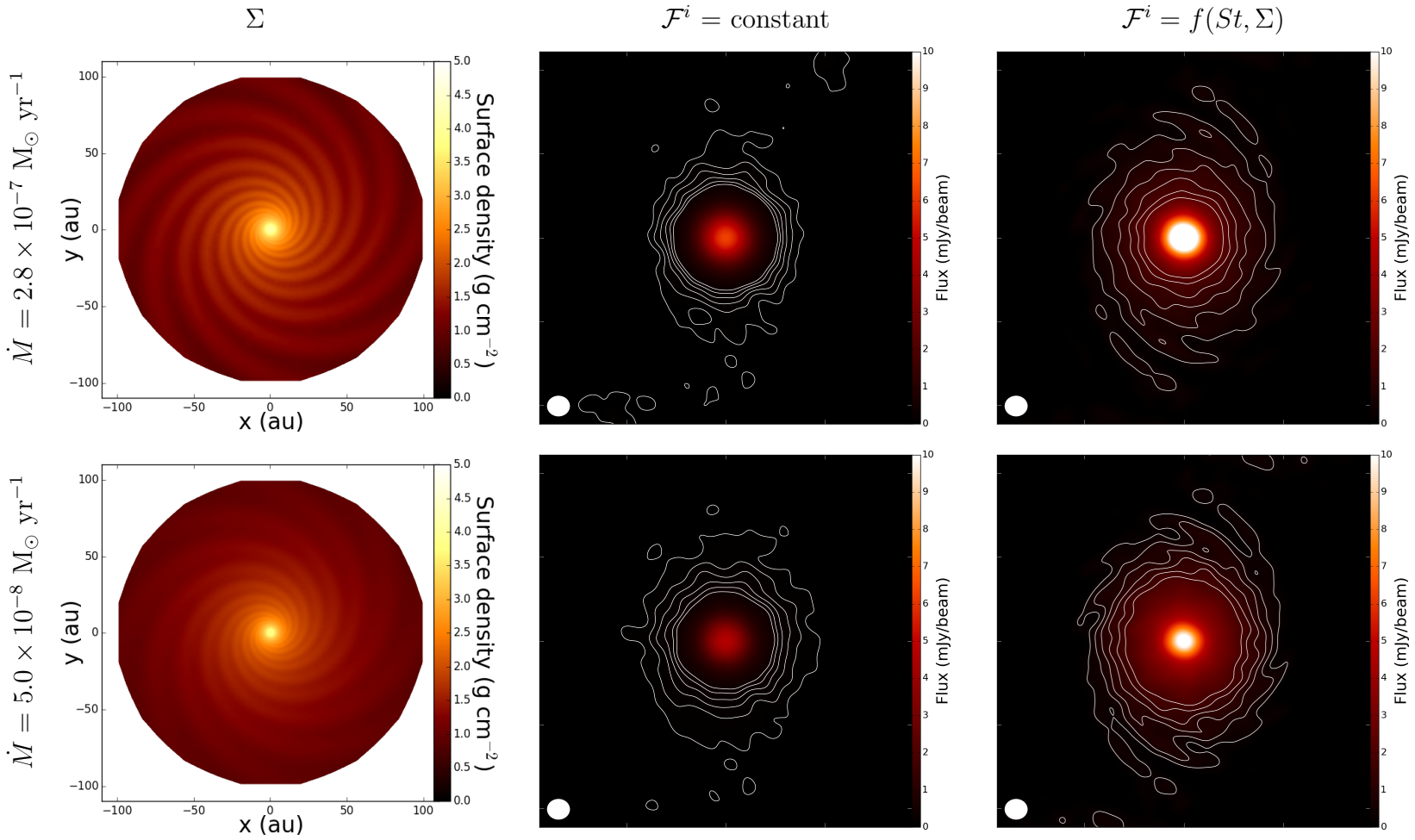


Figure 7.1: Figure showing 4 discs, each disc has 8 spiral arms, the top row has an imposed accretion rate of $\dot{M} = 2.8 \times 10^{-7}$ M$_\odot$ yr^{-1}, bottom row has accretion rate of $\dot{M} = 2.8 \times 10^{-8}$ M$_\odot$ yr^{-1}. In the left hand column, the grain fraction is constant everywhere in the disc. In the right hand column, the total dust is amplified by a factor of 2 inside the spiral arms, and this dust over-density is split between the grains depending upong their Stokes number. All four images were taken using a 1 hour integration with the C40-8.cfg antennalist in the ALMA synthesiser, with a PWV value of 1.262 mm, and an identical beam size, is $0.053'' \times 0.048''$. All contours are the same, at intervals of 3.0, 4.0, 6.0, 9.0, 12.0, 15.0, 18.0, 25.0, 30.0, 50.0 times the RMS in each image. RMS for each image is: top left 18 μJy beam^{-1}, top right 131 μJy beam^{-1}, bottom left 21 μJy beam^{-1}, bottom right 100 μJy beam^{-1}.

We can see immediately that by including our simple analytical prescription for dust trapping, where $\mathcal{F}^i = f(St, \Sigma)$, self-gravitating spiral waves are detectable throughout the disc even for the relatively low accretion rate of $\dot{M} = 5.0 \times 10^{-8}$ $\mathrm{M}_\odot$ yr^{-1}, whereas even for the much higher accretion rate, and therefore disc mass, of the $\dot{M} = 2.8 \times 10^{-7}$ $\mathrm{M}_\odot$ yr^{-1} image with $\mathcal{F}^i$ = constant, it is not possible to detect these spiral features.

To isolate whether this is because we have now allowed $\Sigma_\mathrm{d}/\Sigma_\mathrm{g}$ to peak at 0.02 in spiral arms, or because for grains of $St \sim 1$ we now allow those grains to make up most of the dust population, we ran an additional with imposed $a_\mathrm{max} = 1$ μm. For grain sizes between 0.1 μm and 1 μm, we expect St to be so low, everywhere in the disc, that

$$\mathcal{F}^i \approx \frac{\Sigma_\mathrm{d}}{\Sigma_\mathrm{gas}} \cdot \frac{1}{n_\mathrm{dust}}. \tag{7.27}$$

In this manner, we can isolate the dust trapping of grains of different sizes based upon Stokes number, and *total* dust trapping in the spiral arm. We note, however, that the flux emitted by the dust, for a given frequency, depends on the opacity, κ_ν, of the dust at that frequency, given through (Draine, 2006)

$$F_\nu = \frac{\kappa_\nu M_\mathrm{dust} B_\nu(T_\mathrm{dust})}{D^2}, \tag{7.28}$$

where M_dust is the mass of the emitting particles, B_ν is the blackbody intensity at that frequency, and D is the distance to the disc. Upon examining Figure 7.2, we can see that at $\nu = 220$ GHz, we may expect the flux emitted for $a_\mathrm{max} = 1$ μm to be a factor of ~ 10 lower than for $a_\mathrm{max} = 2000$ μm since the opacity at the two frequencies differs by a factor of 10.

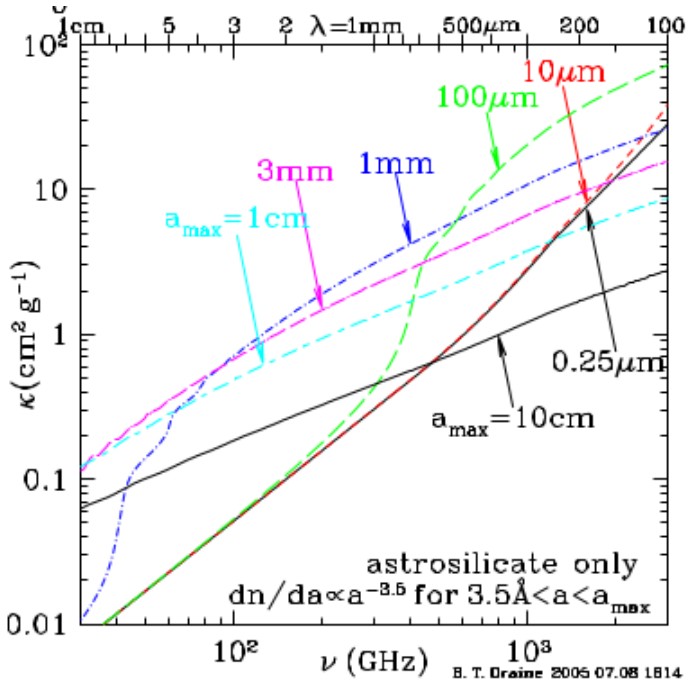


Figure 7.2: This image is Figure 3, taken from Draine (2006). It shows the opacity, κ, of amorphous silicate spheres, with size distribution $dn/da \propto a^{-3.5}$ for different values of $a_{\rm max}$.

However, it is *contrast* that we are examining here, not the total integrated flux. For lower flux objects, we may reasonably expect the signal noise ratio to be lower in the observation, resulting in spiral features that cannot be detected at the 3σ level. This can be compensated for by increasing the integration time on source.

We show our results for this in Figure 7.3, where all three discs have the same physical geometry, $\dot{M} = 2.4 \times 10^{-8}$ M$_\odot$ yr^{-1}, $m = 8$ spiral arms. The radiative transfer calculation for the disc in the left-hand panel was done assuming a maximum grain size of $a_{\rm max} = 2000$ μm, while for the centre and right hand panels a maximum grain size of $a_{\rm max} = 1$ μm was assumed. All discs are imaged

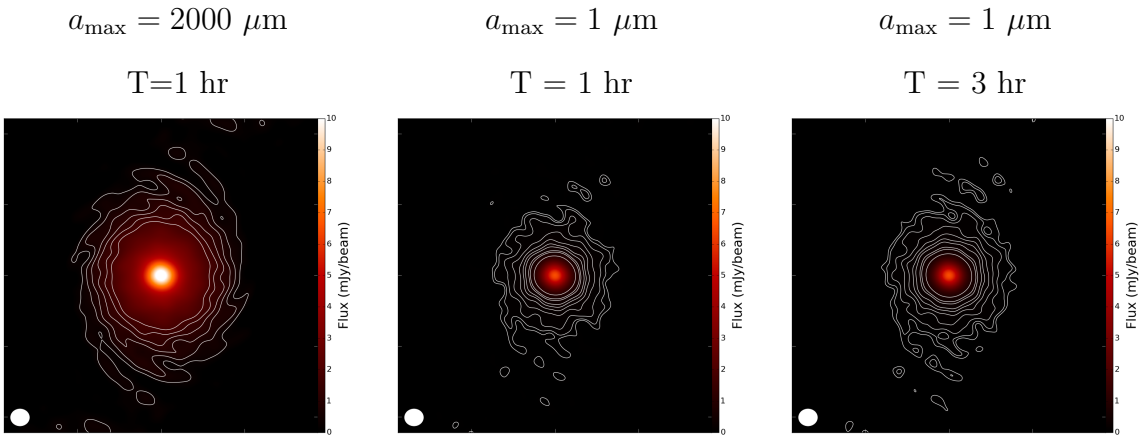


Figure 7.3: All three images have same physical geometry, $\dot{M} = 2.4 \times 10^{-8}$ M$_\odot$ yr^{-1}, $m = 8$ spiral arms. All discs imaged using the C40-8.cfg antennalist, with beam size $0.053'' \times 0.048^{prime\prime}$, shown for reference in the bottom left corner of each panel, central frequency $\nu = 220$ GHz. Left panel: $a_{\rm max} = 2000$ μm, one hour integration, RMS = 100 μJy beam^{-1}. Centre panel: $a_{\rm max} = 1$ μm, RMS=20 μJy beam^{-1}, one hour integration. Right panel: $a_{\rm max} = 1$ μm, RMS=15 μJy beam^{-1}, three hour integration. By comparing the left and centre panel, we can see that for the same exposure time, the extended spiral structure is more difficult to detect if the maximum grain size is 1 μm compared to 2000 μm. This suggests that grain growth may make it easier to detect signatures of disc self-gravity, unless the disc is observed for longer periods of time, as shown in the rightmost panel.

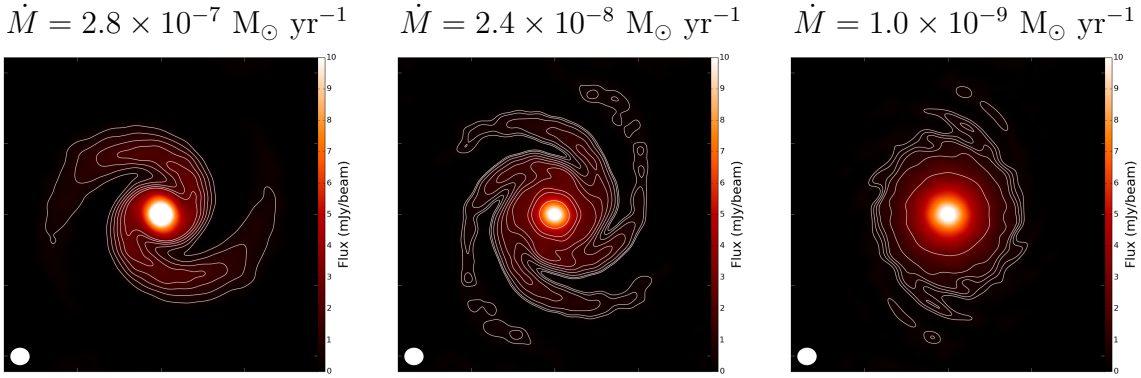


Figure 7.4: Beam size is $0.053'' \times 0.04''$ in all images. Left: $m = 2$, $\dot{M} = 2.8 \times 10^{-7}$ M$_\odot$ yr^{-1}, RMS = 142 μJy beam^{-1}, contours = [3,6,9,12,18]$\times$ RMS. Center: $m = 4$, $\dot{M} = 2.4 \times 10^{-8}$ M$_\odot$, RMS = 98 μJy beam^{-1}, contours = [3,4,6,9,12,15,18,25,30,50]$\times$ RMS. Right: $m = 8$, $\dot{M} = 1.0 \times 10^{-9}$ M$_\odot$ yr^{-1}, RMS= 89 μJy beam^{-1}, contours = [3,4,5,6,9,18]$\times$ RMS. For the last image, we tried several combinations of contours, and this set produced the most visible structure.

at 220 GHz, with a beam size of $0.053'' \times 0.048^{prime\prime}$. Left and centre panel were observed for one hour, right panel was observed for three hours. We can see that spiral structure is more easily detected when $a_{\rm max} = 2000$ μm, unless the disc is observed for much longer periods of time. From this, we may conclude that a disc that has undergone grain growth will be easier to detect signatures of self-gravity in, than a disc with little to no grain growth.

7.4.2 Results for $a_{\rm max} = 2000$ μm

In this section, we run three models with accretion rates $\dot{M} = 2.8 \times 10^{-7}$ M$_\odot$ yr^{-1}, $\dot{M} = 2.4 \times 10^{-8}$ M$_\odot$ yr^{-1} and $\dot{M} = 1.0 \times 10^{-9}$ M$_\odot$ yr^{-1}, with the number of spiral arms as $m = 2$, $m = 4$ and $m = 8$ respectively, based on the empirical relationship $m \sim 1/q$ (Dong et al., 2015b). Our results are shown in Figure 7.4, that shows the three discs, decreasing in accretion rate from left to right, and increasing in spiral

mode from left to right. The beam size is shown in the image bottom left corner, for each image it is identical, $0.053'' \times 0.04''$. Contours are at intervals of the RMS, as given in the figure caption. Naturally, the lower m-mode spirals are better resolved further into the center of the disc, than higher m-mode spirals, given the larger degree of spatial separation that occurs if there are fewer spirals. What is compelling, however, is even at the very low accretion rate of $\dot{M} = 1.0 \times 10^{-9}$ $M_\odot$ yr^{-1}, with $m = 8$, a small amount of non-axisymmetric structure is still detectable. If we compare this to the bottom left image given in Figure 7.1, that has a higher accretion rate and disc mass, yet has no detectable spiral structure, we can conclude that by modelling this dust trapping, we decrease the minimum accretion rate at which signatures of disc self-gravity are detectable. Of course, at such low accretion rates, the gravitational stress in the disc is likely to become dominated by the background MRI. If this is the case, the spirals will begin to be washed out.

7.5 Results for a_{max}=100 cm

We now consider the same three discs as described in section 7.4.2, however, we now perform the radiative transfer calculation with $a_{\rm max} = 100$ cm. We consider such large grain sizes so as to capture the movement of the $St = 1$ boundary from the outer to the inner part of the disc is we increase the grain size. This is shown in Figure 7.5, which shows the stokes number of the $\dot{M} = 2.4 \times 10^{-8}$ $M_\odot$ yr^{-1} disc, plotted for four different grain sizes. We can see that in our analytical model, $St \sim 1$ is located in the inner region of the disc for grains ~ 100 cm in size, then between $50 - 100$ au for grains ~ 1 cm in size, and for grains ~ 1 mm

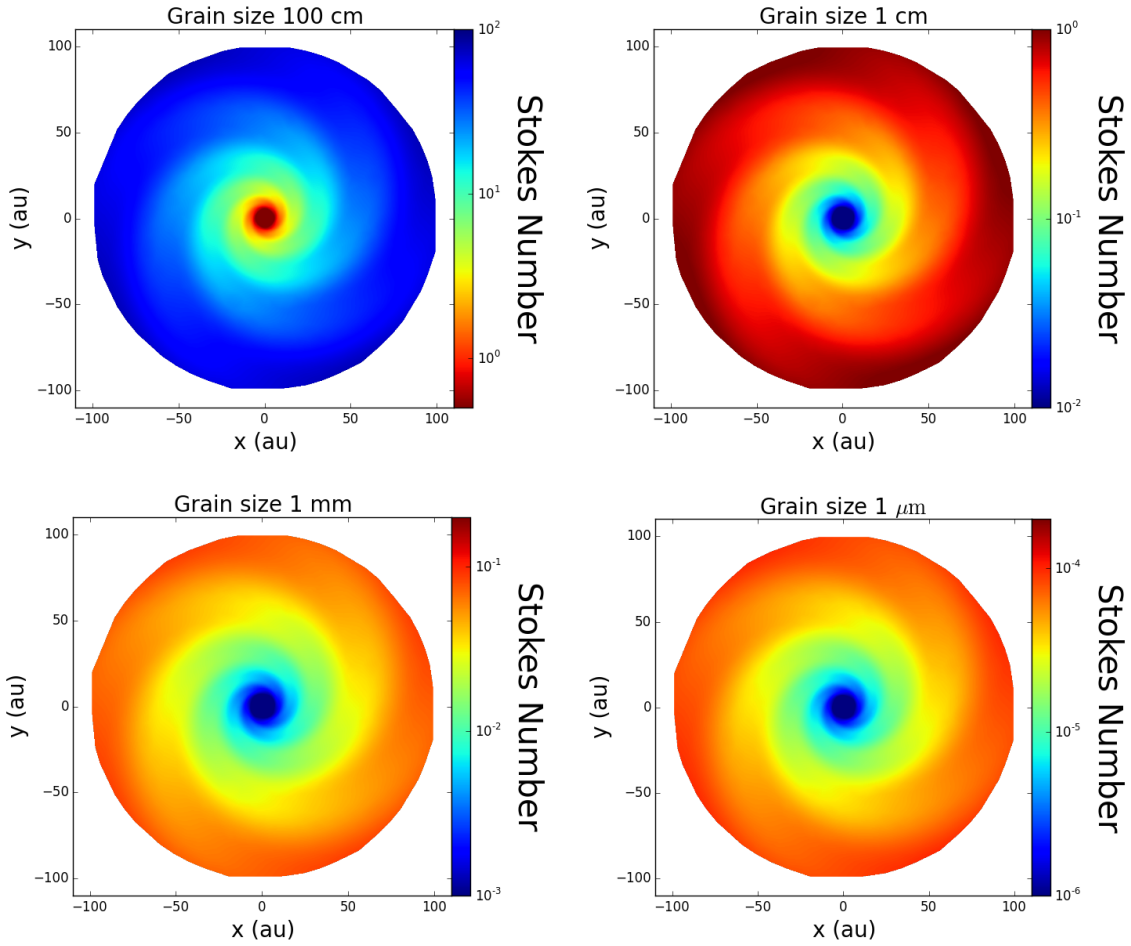


Figure 7.5: Color maps of Stokes Number (St), all in the same disc ($\dot{M} = 2.4 \times 10^{-8}$ $M_\odot$ yr^{-1}), for four different grain sizes. Colour scheme for top row is inverted, both schemes tend towards red for $St \sim 1$.

in size, even at the outermost region in the disc the largest value of St is 0.1. For grains below $\sim$ 1 mm in size, dust is well mixed with the gas and experiences little to no trapping.

We now consider how this is translated to grain abundances in the disc. First, we note that by allowing the grain fraction to grow based on the Stokes number of the grain, we are, essentially, allowing considerable grain growth to happen in the disc. That grain growth is maximised for grain sizes between 1 and 100 cm, since it is for these grain sizes that $St \sim 1$ at least somewhere in the disc, may result in a reduction in total flux from the disc when observed at $\nu = 220$ GHz, since at these grain sizes, $\kappa(\nu) < 1$ (see Figure 7.2). However, in at least some sense, the detection of these spiral arms relies on the *contrast* between the arm and inter-arm region in the disc, rather than the magnitude of the flux in the spiral arms.

This is demonstrated by Figure 7.6, which shows the same three disc geometries as in Figure 7.4, with the only difference being that $a_{\rm max} = 100$ cm now, rather than 2000 μm. Although the flux has reduced by a factor of $\sim 5 - 10$, the contrast is still large enough that the spiral arms can be detected at 3σ even down to $\dot{M} = 1.0 \times 10^{-9}$ M$_\odot$.

We now consider what the distribution of dust, composed of different grain sizes, may look like in our analytical prescription. To this end, we have taken one disc, with $\dot{M} = 2.4 \times 10^{-8}$ M$_\odot$ yr^{-1}, and $m = 4$, and plotted the grain fraction of 6 different dust sizes, from 100 cm to 1 μm, shown in Figure 7.7. The grain fraction is plotted logarithmically, and the St value is overplotted in black contours at

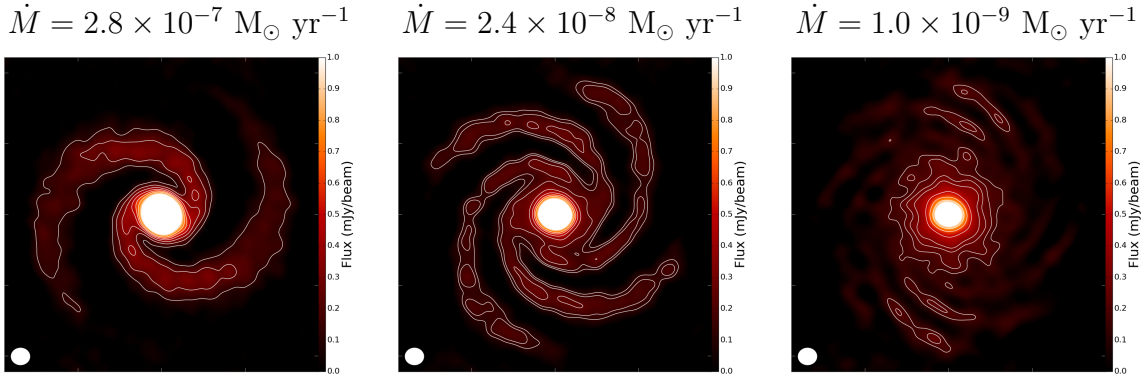


Figure 7.6: Synthesised ALMA images for the same disc geometries as in Figure 7.4, with $a_{\rm max} = 100$ cm. Beam size is $0.053'' \times 0.048''$, shown in bottom left corner of each image, all images produced with 1 hour of integration time with antenna configuration C40-8.cfg, centered on 220 GHz, bandwidth 7.5 GHz, PWV = 1.262 mm. Left: RMS = 39 μJy beam^{-1}, contours=[3.0, 4.0, 5.0, 6.0, 9.0, 18.0]$\times$ RMS Center: RMS = 41 μJy beam^{-1}, contours=[3.0, 4.0, 5.0, 6.0, 9.0, 18.0]$\times$ RMS. Right: RMS = 54 μJy beam^{-1}, contours = [3.0, 4.0, 5.0, 6.0, 9.0, 18.0]$\times$ RMS. Structure is visible that is not outlined by contours in the $\dot{M} = 1.0 \times 10^{-9}$ M$_\odot$ yr^{-1} disc because it falls just below the 3σ threshold.

different points in the disc.

We can see immediately that for the largest grain sizes (100 cm), $St \sim 1$ in the innermost region of the disc, and as the grain size decreases, the location of the most prominent dust-trapping effect moves outwards in the disc. We note that for grain sizes below $\sim$ 1 cm, the Stokes number does not equal 1 anywhere in the disc, which means that the gas/dust mixture is becoming increasingly well mixed. That significant trapping only occurs for grain sizes 1-100 cm is consistent with simulation results such as Rice et al. (2004). Considering the simplicity of our model, that we are able to reproduce the same qualitative results as more sophisticated techniques may imply that we have captured the basic physics that is required, analytically, to reproduce this dust trapping effect.

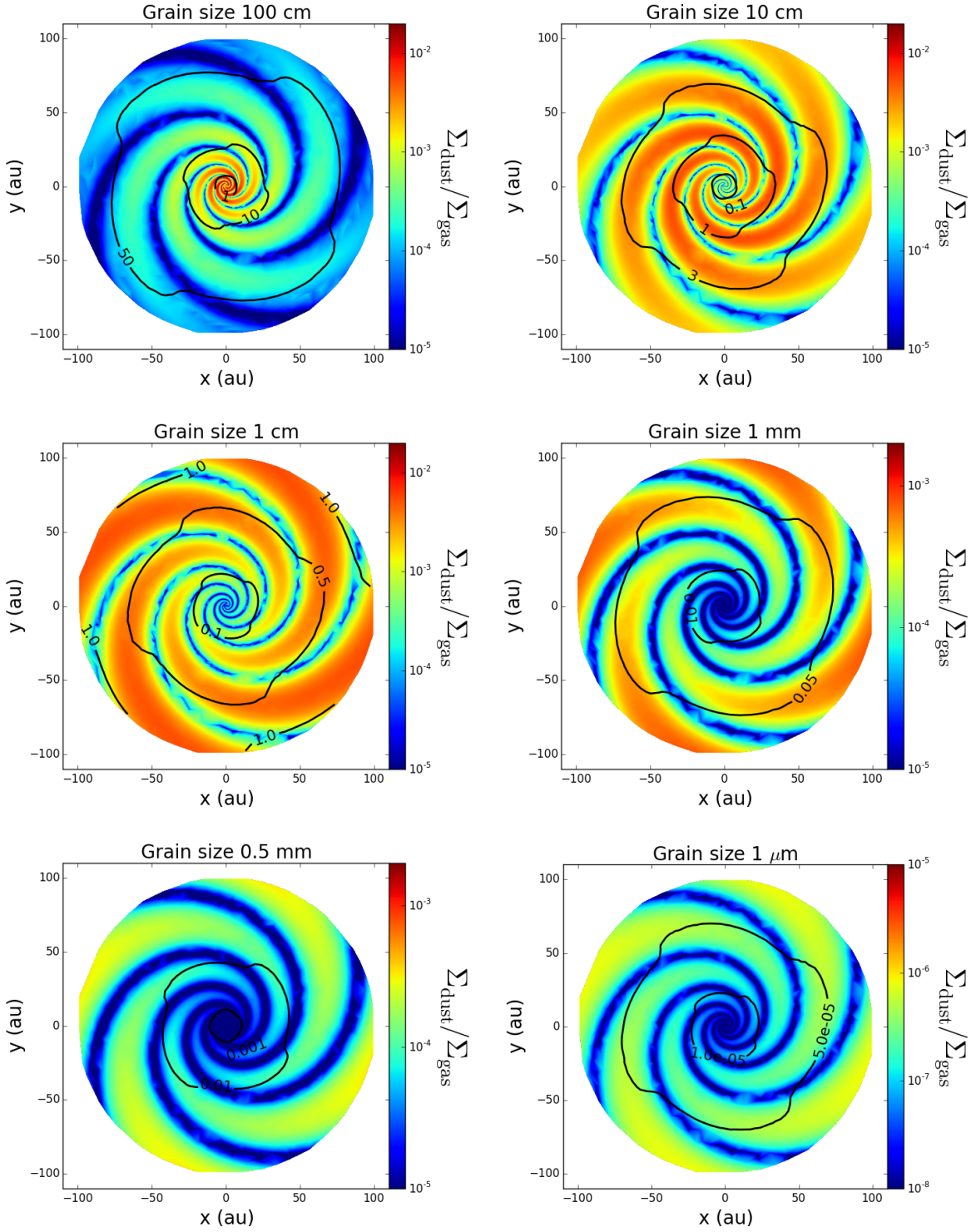


Figure 7.7: Grain fraction maps, for grains of sizes $0.1\,\mu\text{m} < a < 100\,\text{cm}$. Contours of constant stokes number are plotted over the maps in black lines. We see that the $St \sim 1$ for 100 cm in the inner region in the disc, and this contour slowly moves outwards as we decrease in grain size. The dust trapping therefore has a radial dependence on grain size.

7.6 Discussion and Conclusion

In this chapter, we expanded upon the semi-analytical model developed in Chapter 6, to now consider the effects of localised dust enhancement in the spiral arms of the protostellar disc. There is no clear-cut method, in current literature, of describing dust enhancement, without evolving a simulation in time.

We describe an enhancement that is proportional to the radial drift velocity of the particle, such that particles with $St \sim 1$ experience the largest enhancement. We find that by including our dust enhancement prescription, we are able to detect self-gravitating spirals approximately an order of magnitude lower than stated in Chapter 6, i.e. down to $\dot{M} = 5.0 \times 10^{-8}$ M$_\odot$ yr^{-1}. Although we are also able to detect some non-axisymmetry at $\dot{M} = 1.0 \times 10^{-9}$ M$_\odot$ yr^{-1}, realistically, at such a low accretion rate, it is likely that the gravitational stress in the disc will become dominated by the background MRI. If this is the case, the spiral amplitudes will decrease in strength, which decreases the likelihood of them being detected.

By comparing identical ALMA integration times in the left and center panel of Figure 7.3, with $a_{\rm max} = 2000$ μm and $a_{\rm max} = 1$ μm, we have shown that this increased ease of detection is unlikely to occur without significant grain growth having taken place. Therefore, we suggest that if self-gravitating spiral waves are detected, and the dominant m-mode is $m \gtrsim 8$, then grain growth must have taken place if the spiral arms are to be detected.

We have shown that when performing a radiative transfer calculation that uses dust grain sizes in the range from $a_{\rm min} = 0.1$ μm to $a_{\rm max} = 100$ cm, the contrast between the arm and inter-arm regions is still large enough that the ALMA can

resolve the spiral arms at at least 3σ, despite the lower total flux (a factor of $\sim 5-10$) due to the smaller opacity of larger grain sizes at $\nu = 220$ GHz.

By plotting dust fraction maps, we have shown at what radius, for various grain sizes, $St \sim 1$ occurs in the disc. Using this, we may explain why even at accretion rates as low as $\dot{M} = 1.0 \times 10^{-9}$ $M_\odot$ yr^{-1}, we are still able to discern non-axisymmetric structure in the outer regions of the disc. For grains $\sim$ 1 mm in size, the Stokes number does not equal 1 anywhere in the disc. Despite this, there is significant enough contrast in fractional abundance between the arm and interarm regions, especially in the outer part of the disc, that we can detect spiral structure.

Our model is simplistic, and we do not consider the effects of vertical settling in the disc. This, potentially, could be very interesting, since if some of the dust settles to the midplane, it could be masked by the opacity of the upper layers of the disc. If this is the case, then we would, perhaps, expect more difficulty in detecting the spiral structure.

However, since the majority of the continuum opacity is being generated by dust, it is not unreasonable, perhaps, to expect that dust settling may reduce the optical depth at higher altitude in the disc, which would actually allow greater visibility of these spirals. The inclusion of dust settling in our model is therefore a logical next step, and we leave this to future work.

Ultimately, despite the simplicity of our model, we have been able to show that dust enhancement may play an important role in the detection of self-gravitating spiral waves. To do so, dust growth must have happened, since $St \ll 1$ for grain

sizes of ~ 1 μm.

You never really finish your thesis. You sort of just... stop.

Dr. Richard Alexander, talking to CH November 2016

8

Conclusions

It is an incredibly exciting era in the study of protostellar discs, planet formation and exoplanets. At the time of writing, there are 3575 confirmed exoplanets[19], demonstrating the variety of planetary configurations possible, and highlighting that the architecture of our own solar system is not, necessarily, the blueprint for others.

[19]According to the Extrasolar Planets Encyclopaedia online catalogue, available here: http://exoplanet.eu/catalog/.

8.1 Thesis Summary

Exoplanets form in protostellar accretion discs; as we are now into the ALMA era of observational astronomy, we are availed of a panoply of unprecedentedly high resolution images of these discs. These observations are crucial in verifying theoretical models of planet formation.

In this thesis, I have discussed my research into fragmenting and non-fragmenting self-gravitating protostellar discs. This research has been conducted using Smoothed Particle Hydrodynamics (SPH) simulations, semi-analytic models, radiative transfer calculations and synthetic imaging. I ran a suite of SPH simulations, using different implementations of SPH, together with different artificial viscosity prescriptions, in order to investigate the apparent lack of convergence of the fragmentation boundary with increasing resolution. My investigation did not lead to a decisive conclusion for the fragmentation problem, but in hindsight this was unsurprising, since it was found (while the work presented in this thesis was underway) that by not using smoothed cooling, one actually introduces a mismatch in length scales.

What is interesting, however, is that discs with identical initial conditions, if evolved using different implementations of SPH, would take remarkably different paths. The amount of fragmentation may be greatly altered, or fragmentation may be altered altogether. This should, therefore, inform us that the choice of exact implementation of SPH is important, and should be treated carefully when conducting numerical tests.

I have developed a new algorithm to identify fragments in SPH simulations,

the density derivative search (DDS), and compared it to an already existing algorithm, `CLUMPFIND`. I find that these two algorithms are complementary, the DDS is able to detect all fragments from an earlier stage, regardless of whether they survive the duration of the simulation, while the `CLUMPFIND` algorithm does not find clumps that are ultimately destroyed. By using both, one can therefore make predictions about which clumps are likely to be destroyed.

I then use the DDS algorithm, together with a standard merger-tree algorithm, to track and analyse fragments formed in 9 simulations of fragmenting protostellar discs. I have compared the results of these simulations to gravitational instability (GI) population synthesis models, and find that at early times, the DDS algorithm underestimates fragment mass, however is quite robust a few orbital periods later. In light of this, I provide two distribution fits, one to fragment eccentricity and one to fragment inclination, that I suggest may be useful in the development of future GI population synthesis models.

Additionally, I find that polytropes may not be the best representation of fragments in population synthesis models, since the fragments in the SPH simulations are typically an order of magnitude denser in the very center. This could significantly alter orbital parameters that are sensitive to the mass distribution of the fragments. Furthermore, I also find that the environment in which the fragment forms may play a critical role in its ultimate fate. Spiral arms may shock fragments, increasing their internal energy, which may cause dust to sublimate before the fragments form solid cores. If this is frequently true, then planet formation through tidal downsizing may be very difficult. In a similar vein, I find that fragment-fragment interactions may increase the number of fragments

that survive in a simulation, since if they are scattered out to large radii, they are less likely to be tidally destroyed by the central star. This could, potentially, have important implications for the GI planet formation hypothesis, if fewer fragments are destroyed than originally thought.

If a protostellar disc does not cool rapidly enough, it will settle into a quasi-steady state. I have taken a pre-existing, 1D analytical model of a quasi-steady, self-gravitating protostellar disc, and converted into a model that can describe the 3D structure of such a quasi-steady, self-gravitating disc. I perform radiative transfer calculations on discs that satisfy a range of accretion rates, and therefore disc masses, and find, by generating synthetic images, that there is a very small region of parameter space where the stresses in the disc are large enough to generate spiral arms that are detectable by an instrument such as ALMA, but not so strong as to cause the disc to fragment. However, by including a simple analytical prescription for dust enhancement inside these spiral arms, this region of parameter space broadens considerably, and we are able to detect self-gravitating spiral arms down to an accretion rate an order of magnitude lower than previously thought. I have shown that this is only possible, for a given instrument and integration time, if grain growth has happened. I therefore conclude that if self-gravitating spiral arms are detected, it is likely that at least some grain growth has taken place.

8.2 Future Prospects

Modelling, whether fully numerical or semi-analytical, is a powerful tool in an astronomer's arsenal. For processes that are well understood, or rather,

considered well understood, semi-analytical models allow us to examine a large parameter space without the computational cost of fully numerical experiments. I have shown in this thesis that making some very simple assumptions about the physical distribution of dust in a semi-analytical model alters the parameter space in which signatures of disc self-gravity would be detected by ALMA.

With this in mind, a clear area for future development is in the analytical understanding of the behaviour of dust. Although the equations of motion that describe the time behaviour of dust have been known since the 1970's (Weidenschilling, 1977; Whipple, 1972), there is, currently, no completely self-consistent prescription for the steady state dust profile in a disc.

All of the SPH simulations in this thesis considered only the dynamics of the gas component of the disc. When post-processed with MCRT to produce synthetic images, it was assumed that gas and dust are well mixed. However, we know that, in reality, grains with $St \sim 1$ will experience significant enhancement at pressure maxima. Ideally, multiple grain species should be modelled simultaneously during the dynamical evolution of the disc, and this can then be post-processed using MCRT. One, as of yet, unanswered question is how, or if, dusty discs will fragment to form gravitationally bound objects. Orbital parameters of objects are, in general, sensitive to their radial distribution of mass. It would be very interesting to see how physical processes, such as dust settling (if it occurs) inside these fragments, affects orbital parameters.

More generally, simulating fragmenting dusty discs could have some interesting implications for the *tidal downsizing hypothesis* (Nayakshin, 2010a, 2015a,d;

Nayakshin & Fletcher, 2015) theory of planet formation. In this scenario, a gas clump of ~ 1 M_J forms ~ 100 au from the central star, through gravitational instability. The clump then migrates inwards due to disc torques, all the while grain sedimentation and the gravitational instability of the solid component of the clump is acting to form a solid core. The fragment is either destroyed by tides from the central star, if gravitational contraction has not happened rapidly enough, or, if contraction is faster than migration, the clump survives and becomes a gas-giant planet.

However, the population synthesis models of Forgan & Rice (2013b) have shown that around half of the initially formed clumps are destroyed, although this may be significantly lower if stochastic processes, such as fragment-fragment scattering, are considered (see Chapter 5). Furthermore, the majority ($\sim 90\%$) of clumps that do form are above the Deuterium burning limit ($\gtrsim 13$ M_J). The population synthesis model of Forgan & Rice (2013b) does not include accretion onto the clumps from the surrounding disc, so the masses obtained are, in fact, lower limits to the expectation values of fragment masses. However, there are several potential mechanisms by which growth may be limited. For example, Stamatellos (2015) has shown that accretion-powered radiative feedback from the planet significantly limits its mass growth in the early stages of disc evolution. However, this was coupled with gap opening that essentially halted further inwards migration. Later work (Nayakshin, 2015b,c) has shown that *pebble accretion* may lower the destruction rate of the clumps, by altering the opacity and therefore increasing the gravitational collapse rate.

From the above, it is easy to see that hydrodynamic simulations that fully

capture radiative feedback from potential protoplanets to the surrounding gas, that also include all the necessary physics to correctly model dust sedimentation in fragments, would greatly improve our current understanding of planet formation theory, particularly for the tidal downsizing hypothesis. If, by modelling pebble accretion, the destruction rate of the clumps is decreased, and, if including radiative feedback from the fragments to the gas lowers the minimum mass of fragments, then the tidal downsizing hypothesis may, possibly, capture a good portion of the physics involved in planet formation. How, statistically, a sample drawn from a suite of such simulations would compare to known exoplanets would require careful attention. More generally, a full statistical comparison between the known exoplanets, and population synthesis models for gravitational instability and core accretion is probably desirable, and may help us better constrain the parameter space in which these regimes operate.

8.3 Closing Remarks

This is an exciting time to be an astronomer who is interested in protostellar discs and exoplanets. With ALMA, and the advent of the James Webb Space Telescope (JWST) and the European Extremely Large Telescope (E-ELT), it is likely that the next decade will be filled with first time discoveries, characterisations and developing theories. We are part of an exciting, diverse research community, faced with many grand challenges (see, e.g. Haworth et al. 2016). These challenges are numerous, difficult, and computationally expensive, all of which is daunting. However, by knowing what they are, the exoplanet and protostellar disc community is able to collaboratively focus their research efforts. This thesis

presents work that contributes to mankind's understanding of the role of disc self-gravity in planet formation, and, ultimately, the origin of the Solar System, and its place in the Universe.

A

Spiral Structure and the Derivation of the Toomre Parameter From the Dispersion Relation

Up until the 1950s, astronomers thought galactic spiral structure was due to the interaction of interstellar gas with the magnetic field of the galaxy. However, it was eventually realised that the strength of magnetic fields in galaxies was far too weak to be the cause of this structure.

It was Lindblad who realised that spiral structure originates from the interaction between the orbits and gravitational forces of the disc. However, it was not until Lin and Shu's work whilst they were at MIT that the numerical framework was in place to study this further. They applied the physics of wave mechanics to study spiral density waves, realising that periodic compression and rarefaction of the disc surface density, analogous to wave propagation over the surface of the ocean, would lead to a spiral density wave. Their ideas are combined into what is known as the *Lin-Shu hypothesis*, that spiral structure is a stationary density wave, unchanged except for an overall rotation.

Although it is certainly true that this spiral structure is long lasting, it was later realised that it was transient in nature. The work of Goldreich and Lynden-Bell demonstrated that differentially rotating, self-gravitating discs are responsive to even small disturbances, and therefore spiral structure may be transient in nature.

In order for us to proceed analytically, we must make some assumptions about the nature of our system.

1. That any perturbations to the system are sufficiently small that they may be described by linear perturbation theory.

2. We assume that the spirals are sufficiently tightly wound (known as the WKB approximation) that the long range gravitational force may be neglected. In other words, the response of the matter to gravitational perturbations is local.

3. That the spiral structure is quasi-steady. It remains relatively unchanged

over many rotation periods.

Points 1 and 3 need little further explanation, however, the WKB approximation is a little more involved, so we elucidate upon it now. For the most part, the derivations in this appendix follow those found in Binney & Tremaine (2008).

A.1 The WKB Approximation

The WKB (Wentzel-Kramers-Brillouin) approximation, also known as the *tight-winding* approximation, removes long-range coupling effects by assuming the waves are tightly wound. Waves can only be described as tightly wound is a certain condition is satisfied, and we explain this now.

If we consider the equation of a curve on the surface of a disc, $\phi + g(R,t) =$ constant, where ϕ is the azimuthal angle, R is the distance from the center of the disc, and t is time, we can describe a system with m-fold rotational symmetry as

$$m\phi + f(R,t) = \text{constant} \pmod{2\pi}, \tag{A.1}$$

so long as $m > 0$. Here,

$$f(R,t) \equiv mg(R,t) \tag{A.2}$$

is known as the *shape function*. We now introduce the radial wavenumber,

$$k(R,t) \equiv \frac{\partial f(R,t)}{\partial R}, \tag{A.3}$$

and the pitch angle, α, which is determined through

$$\cot\alpha = \left| R\frac{\partial\phi}{\partial R} \right|, \tag{A.4}$$

where α is the angle, at radius R, between the tangent to the spiral arm, and the tangent to the circle, of radius R. The partial derivative is evaluated along the curve using equation A.1. Therefore,

$$\cot\alpha = \left| \frac{kR}{m} \right|. \tag{A.5}$$

Now, we lay down the assumption that the waves are tightly wound. The radial separation between adjacent arms is ΔR, where we have

$$|f(R+\Delta R, t) - f(R,t)| = 2\pi, \tag{A.6}$$

however, so long as they waves are tightly wound, i.e. ΔR is sufficiently small, then we can rewrite equation A.6 as

$$|f(R,t) + \left(\frac{\partial f}{\partial R}\Delta R\right) - f(R,t)| = 2\pi. \tag{A.7}$$

Seeing that the first and last terms of equation A.7 cancel, we can make use of equations A.4 and A.5 to find

$$\Delta R = \frac{2\pi}{|k|} = \frac{2\pi R}{m}\tan\alpha, \tag{A.8}$$

so the radial wavelength is therefore $\frac{2\pi}{|k|}$ in the WKB approximation. The WKB

approximation requires $|kR| \gg 1$, or, equivalently, $\cot\alpha \gg 1$.

A.2 The Dispersion Relation

So long as the WKB approximation is valid, then for an infinitesimally thin disc responding to a tightly wound spiral perturbation, we can determine the *dispersion relation*. Broadly speaking, the dispersion relation is the function $\omega(k)$, where omega is the frequency (in this case, the angular frequency), and k is the wavenumber, i.e. the number of waves per unit distance. If the speed of propagation, c, is constant, then the dispersion relation is simply

$$\omega(k) = ck. \tag{A.9}$$

The dispersion relation, therefore, describes the effects of dispersion in a medium (when different frequencies have different propagation speeds) on the properties of a wave travelling in that medium.

We begin with Euler's equations for the radial and azimuthal velocity components in a disc:

$$\frac{\partial v_r}{\partial t} + v_r\frac{\partial v_r}{\partial t} + \frac{v_\phi}{r}\frac{\partial v_r}{\partial \phi} - \frac{v_\phi^2}{r} = -\frac{\partial \Phi}{\partial r} - \frac{1}{\Sigma}\frac{\partial P}{\partial R} \tag{A.10}$$

$$\frac{\partial v_\phi}{\partial t} + v_r\frac{\partial v_\phi}{\partial t} + \frac{v_\phi}{r}\frac{\partial v_\phi}{\partial \phi} + \frac{v_r v_\phi}{r} = -\frac{1}{R}\frac{\partial \Phi}{\partial \phi} - \frac{1}{R\Sigma}\frac{\partial P}{\partial \phi}. \tag{A.11}$$

We have replaced ρ with Σ, since we are now dealing with a two-dimensional disc.

Now, we assume an equation of state of the form

$$P = K\Sigma^{\gamma}, \tag{A.12}$$

and note that the specific enthalpy, h, is given by

$$h = \int \frac{dP}{\rho}. \tag{A.13}$$

Using equation A.12, equation A.13 becomes

$$h = \frac{\gamma}{\gamma - 1} K\Sigma^{\gamma - 1}. \tag{A.14}$$

This can be used to simplify the right hand side of equation A.10, such that

$$-\frac{\partial \Phi}{\partial r} - \gamma K\Sigma^{\gamma-2}\frac{\partial \Sigma}{\partial R} = \frac{\partial}{\partial R}(\Phi + h), \tag{A.15}$$

and similarly for equation A.11. If we now assume that the spiral wave is simply a small perturbation on an axisymmetric disc, then we can rewrite our variables as linear perturbations, i.e.

$$\begin{aligned} v_R &= V_{R0} + \eta v_{R1} \\ v_\phi &= V_{\phi 0} + \eta v_{\phi 1} \\ h &= h_0 + \eta h_1 \\ \Sigma &= \Sigma_0 + \eta \Sigma_1 \\ \Phi &= \Phi_0 + \eta \Phi_1, \end{aligned} \tag{A.16}$$

where $\eta \ll 1$, and variables with subscript 1 are the same order of magnitude as variables with subscript 0. The variables with subscript 0 represent the unperturbed disc, and since we expect the unperturbed disc to be radially stable and axisymmetric, then $v_{r0} = 0$, $\frac{\partial \Phi_0}{\partial \phi} = 0$, $\frac{\partial P_0}{\partial \phi} = 0$ and $\frac{\partial h_0}{\partial \phi} = 0$. Therefore, Euler's equations for the unperturbed disc are now

$$\frac{v_{\phi 0}^2}{R} = \frac{d}{dR}(\Phi_0 + h_0) = \frac{d\Phi_0}{dR} + v_s^2 \frac{d}{dR} \ln \Sigma_0, \tag{A.17}$$

where $\frac{v_{\phi 0}^2}{R}$ is the centripetal acceleration, $\frac{d\Phi_0}{dR}$ is the gravitational potential and $v_s^2 \frac{d}{dR} \ln \Sigma_0$ is the pressure force per unit mass. If the sound speed is much less than the rotation speed, then $v_s^2 \frac{d}{dR} \ln \Sigma_0$ can be neglected in equation A.17, and we can write

$$v_\phi 0 \approx \sqrt{\left(R \frac{d\Phi_0}{dR} \right)} = R\Omega(R), \tag{A.18}$$

where $\Omega(R)$ is the circular frequency as usual. We are now in a position to rewrite the perturbed Euler equations of motion, equation A.10 and equation A.11, using equations A.16, so we now have

$$\frac{\partial v_{R1}}{\partial t} + \Omega \frac{\partial v_{R1}}{\partial \phi} - 2\Omega v_{\phi 1} = -\frac{\partial}{\partial R}(\Phi_1 + h_1) \tag{A.19}$$

$$\frac{\partial v_{\phi 1}}{\partial t} + v_{R1} \frac{d(\Omega R)}{dR} + v_{R1}\Omega + \Omega \frac{\partial v_{\phi 1}}{\partial \phi} = -\frac{1}{R} \frac{\partial}{\partial \phi}(\Phi_1 + h_1). \tag{A.20}$$

It is convenient to introduce a new variable, $B(R)$, defined as

$$B(R) = -\left(\Omega + \frac{1}{2} R \frac{d\Omega}{dR} \right), \tag{A.21}$$

and related to the epicyclic frequency by

$$\kappa^2 = -4B\Omega, \tag{A.22}$$

so now the second and third terms on the left hand side of equation can be rewritten as

$$v_{R1}\frac{d(\Omega R)}{dR} + v_{R1}\Omega = -2B(R). \tag{A.23}$$

In a Keplerian disc, we note that $\kappa = \Omega$. In order to proceed to the dispersion relation, we assume a wave solution to equations A.19 and A.20, such that

$$v_{R1} = \mathrm{Re}[v_{Ra}(R)e^{i(m\phi-\omega t)}], \tag{A.24}$$

and similarly for $v_{\phi 1}$, Φ_1, h_1 and Σ_1, where m is an integer, greater than zero, indicating m-fold rotational symmetry (i.e., the number of spiral arms). Substituting these solutions into equations A.19 and A.20, gives, for v_{Ra} and $v_{\phi a}$:

$$v_{Ra} = \frac{i}{\kappa^2 - (\omega - m\Omega)^2}\left[(\omega - m\Omega)\frac{d}{dR}(\Phi_a + h_a) - \frac{2m\Omega}{R}(\Phi_a + h_a)\right] \tag{A.25}$$

$$v_{\Phi a}(R) = \frac{1}{\kappa^2 - (\omega - m\Omega)^2}\left[- 2B\frac{d}{dr}(\Phi_a + h_a) + \frac{m(m\Omega - \omega)}{r}(\Phi_a + h_a)\right]. \tag{A.26}$$

The linearised enthalpy is given by

$$h_a = \gamma K \Sigma_0^{\gamma-2}\Sigma_a = \frac{v_s^2\Sigma_a}{\Sigma_0}. \tag{A.27}$$

The perturbed surface density is related to the perturbed velocities by the

continuity equation, which gives

$$\frac{\partial \Sigma_1}{\partial t} + \Omega \frac{\partial \Sigma_1}{\partial \phi} + \frac{1}{R}\frac{\partial}{\partial R}(R v_{R1} \Sigma_0) + \frac{\Sigma_0}{R}\frac{\partial v_{\phi 1}}{\partial \phi} = 0. \tag{A.28}$$

We can then make use of our wave solutions given in A.24, and similarly for the rest of our variables, so we now have

$$-i(\omega - m\Omega)\Sigma_a + \frac{1}{R}\frac{d}{dR}(R v_{Ra} \Sigma_0) + \frac{im\Sigma_0}{R} v_{\phi a} = 0. \tag{A.29}$$

Equations A.25, A.26, A.27 and A.29 provide four constraints on the five variables Σ_a, v_{Ra}, $v_{\phi a}$, h_a and Φ_a. The fifth is obtained by linking the gravitational potential to the surface density, using Poisson's equation for the surface density. We proceed first by stating that the gravitational potential of a tightly wound spiral is given by

$$\Phi_a(R) = F(r) e^{if(R)} = F(r) e^{i \int k \, dR}, \tag{A.30}$$

where k is the already-defined radial wavenumber, and $|kR| \gg 1$. Poisson's equation, relating potential and surface density, is given by

$$\Phi_1(R, \phi, t) = -\frac{2\pi \mathrm{G}}{|k|} \Sigma_1(R, \phi, t), \tag{A.31}$$

where

$$\Sigma_1(R, \phi, t) = H(R, t) e^{i(m\phi) + f(R,t)}, \tag{A.32}$$

where $f(R, t)$ is the shape function, and $H(R, t)$ is a slowly varying function of radius that gives the amplitude of the spiral. This expression has a fractional error $\mathcal{O}(|kR|^{-1})$. Now, the disc response to the perturbation, Σ_a, and the enthalpy

response, h_a (through equation A.27), share with Φ_a the complex exponential $e^{if(R)}$. Therefore, terms that involve $\frac{\Phi_a+h_a}{R}$ are smaller than those proportional to $\frac{d\Phi_a+h_a}{dR}$ by a factor of kR, so they can be neglected while maintaining the same order of error.

Similarly, we can write

$$\frac{d(\Phi_a + h_a)}{dR} = ik(\Phi_a + h_a) \tag{A.33}$$

with no additional increase in error. Equations A.25 and A.26 are now

$$v_{Ra} = -\frac{(\omega - m\Omega)k(\Phi_a + h_a)}{\kappa^2 - (m\Omega - \omega)^2} v_{\phi a} = \frac{2ikB(\Phi_a + h_a)}{\kappa^2 - (m\Omega - \omega)^2}, \tag{A.34}$$

then we can also simplify the continuity equation:

$$-(\omega - m\Omega)\Sigma_a + k\Sigma_0 v_{Ra} = 0. \tag{A.35}$$

Eliminating v_{Ra} using equation A.34, h_a using A.27 and Φ_a using A.31, and some rearrangement, finally gives us the *dispersion relation*

$$(\omega - m\Omega)^2 = \kappa^2 - 2\pi \mathrm{G}\Sigma|k| + v_s^2 k^2. \tag{A.36}$$

For the case of Keplerian disc, using $\omega = m\Omega_p$, where Ω_p is the pattern speed, we may instead write

$$m^2(\Omega - \Omega_p)^2 = \kappa^2 - 2\pi \mathrm{G}\Sigma|k| + v_s^2 k^2. \tag{A.37}$$

A.3 Derivation of the Toomre Parameter

The Toomre parameter (Toomre, 1964), is, essentially, a measure of local stability to axisymmetric perturbations in a self-gravitating disc. Perturbations may grow under their own self-gravity, which destabilises the disc. In the case of a differentially rotating disc, these perturbations grow into spiral density waves. We now show under what conditions a spiral density wave may grow and sustain its amplitude in the disc.

First, we note that the dispersion relation in equation A.36 holds for axisymmetric perturbations (i.e., $m = 0$), so long as $|kR| \gg 1$. We begin, first of all, with the dispersion relation for axisymmetric disturbances, given by

$$\omega^2 = \kappa^2 - 2\pi \mathrm{G}\Sigma|k| + v_s^2 k^2. \tag{A.38}$$

All quantities on the right hand side of equation A.38 are real, so ω^2 must also be real. If $\omega^2 > 0$, then ω is real and the disc is stable. However, if $\omega^2 < 0$, then ω is imaginary, for example, if $\omega 2 = -p^2$ then $\omega = \pm ip$, and then $e^{-i\omega t} = e^{\pm pt}$, so for $\omega^2 < 0$, there will always be a perturbation that grows exponentially with time.

Now, we have worked out that $\omega^2 > 0$ is always stable, and $\omega^2 < 0$ is always unstable. Therefore, the line of instability is at $\omega = 0$,so

$$\kappa^2 - 2\pi \mathrm{G}\Sigma|k| + v_s^2 k^2 = 0. \tag{A.39}$$

This is only stable if there are no solutions for any positive $|k|$, i.e. the roots

of equation A.39 must be imaginary. The discriminant[20] of equation A.39 must then be

$$(2\pi G\Sigma)^{\frac{1}{2}} - (2v_s\kappa)^2 < 0, \tag{A.40}$$

rearranging and taking the square root gives us the *Toomre Stability Criterion*,

$$Q \equiv \frac{v_s\kappa}{\pi G\Sigma} > 1, \tag{A.41}$$

which, for a keplerian disc (where $\kappa = \Omega$) is simply

$$Q \equiv \frac{v_s\Omega}{\pi G\Sigma} > 1. \tag{A.42}$$

[20]The discriminant,D, of a quadratic equation $ax^2 + bx + c = 0$ is $b^2 - 4ac$. If $D > 0$, then there are two real solutions. If $D = 0$, there is one real solution. If $D < 0$, there are two complex solutions.

Bibliography

ALMA Partnership et al., 2015, ApJ, 808, L3

Andrews S. M., Wilner D. J., Espaillat C., Hughes A. M., Dullemond C. P., McClure M. K., Qi C., Brown J. M., 2011a, ApJ, 732, 42

Andrews S. M., Wilner D. J., Espaillat C., Hughes A. M., Dullemond C. P., McClure M. K., Qi C., Brown J. M., 2011b, ApJ, 732, 42

Andrews S. M., et al., 2016, ApJ, 820, L40

Armitage P. J., 2011, ARA&A, 49, 195

Armitage P. J., Livio M., Pringle J. E., 2001, MNRAS, 324, 705

Bai X.-N., Goodman J., 2009, ApJ, 701, 737

Baines D., Oudmaijer R. D., Porter J. M., Pozzo M., 2006, MNRAS, 367, 737

Balbus S. A., Hawley J. F., 1991, ApJ, 376, 214

Barnes J., Hut P., 1986, Nature, 324, 446

Barnes J. E., Hut P., 1989, ApJS, 70, 389

Bate M. R., Bonnell I. A., Price N. M., 1995, MNRAS, 277, 362

Batygin K., Bodenheimer P. H., Laughlin G. P., 2016, ApJ, 829, 114

Bell K. R., Lin D. N. C., 1994, ApJ, 427, 987

Benisty M., et al., 2015a, A&A, 578, L6

Benisty M., et al., 2015b, A&A, 578, L6

Berger M. J., Colella P., 1989, Journal of Computational Physics, 82, 64

Berger M., Oliger J., 1984, Journal of Computational Physics, 53, 484

Biller B., et al., 2012, ApJL, 753, L38

Binney J., Tremaine S., 2008, Galactic Dynamics: Second Edition. Princeton University Press

Bodenheimer P., Yorke H. W., Rozyczka M., Tohline J. E., 1990, ApJ, 355, 651

Bodenheimer P., Hubickyj O., Lissauer J. J., 2000, Icarus, 143, 2

Boley A. C., Hayfield T., Mayer L., Durisen R. H., 2010, Icarus, 207, 509

Boley A. C., Helled R., Payne M. J., 2011, ApJ, 735, 30

Boley A. C., Granados Contreras A. P., Gladman B., 2016, ApJ, 817, L17

Bonnell I. A., Rice W. K. M., 2008, Science, 321, 1060

Boss A. P., 1997, Science, 276, 1836

Boss A. P., 1998, ApJ, 503, 923

Boss A. P., 2000, ApJ, 536, L101

Boss A. P., 2002, ApJ, 567, L149

Cai K., Durisen R. H., Michael S., Boley A. C., Mejía A. C., Pickett M. K., D'Alessio P., 2005, in Protostars and Planets V Posters. p. 8155 (arXiv:astro-ph/0508354)

Cameron A. G. W., 1978, Moon and Planets, 18, 5

Casassus S., et al., 2013, Nature, 493, 191

Cha S.-H., Whitworth A. P., 2003, MNRAS, 340, 73

Chandrasekhar S., 1947, ApJ, 106, 145

Chow E., Monaghan J. J., 1997, Journal of Computational Physics, 134, 296

Christiaens V., Casassus S., Perez S., van der Plas G., Ménard F., 2014, ApJL, 785, L12

Clarke C. J., 2009, MNRAS, 396, 1066

Cleary P. W., Monaghan J. J., 1999, JCAP, 148, 227

Close L. M., et al., 2014, ApJL, 781, L30

Cloutier R., Lin M.-K., 2013, MNRAS, 434, 621

Cossins P., Lodato G., Clarke C. J., 2009, MNRAS, 393, 1157

Cossins P., Lodato G., Testi L., 2010, MNRAS, 407, 181

Courant R., Friedrichs K., Lewy H., 1928, "Mathematische Annalen", 100, 32

D'Angelo G., Lubow S. H., 2010, ApJ, 724, 730

Damjanov I., Jayawardhana R., Scholz A., Ahmic M., Nguyen D. C., Brandeker A., van Kerkwijk M. H., 2007, ApJ, 670, 1337

Dipierro G., Lodato G., Testi L., de Gregorio Monsalvo I., 2014, MNRAS, 444, 1919

Dipierro G., Pinilla P., Lodato G., Testi L., 2015, MNRAS, 451, 974

Dong R., Zhu Z., Rafikov R. R., Stone J. M., 2015a, ApJL, 809, L5

Dong R., Hall C., Rice K., Chiang E., 2015b, ApJ, 812, L32

Douglas T. A., Caselli P., Ilee J. D., Boley A. C., Hartquist T. W., Durisen R. H., Rawlings J. M. C., 2013, MNRAS, 433, 2064

Draine B. T., 2006, ApJ, 636, 1114

Draine B. T., Lee H. M., 1984, ApJ, 285, 89

Dunham M. M., et al., 2014, preprint, (arXiv:1401.1809)

Durisen R. H., Boss A. P., Mayer L., Nelson A. F., Quinn T., Rice W. K. M., 2007, Protostars and Planets V, pp 607–622

Fang J., Margot J.-L., 2013, ApJ, 767, 115

Figueira P., et al., 2012, A&A, 541, A139

Fischer D. A., Valenti J., 2005, ApJ, 622, 1102

Forgan D., Rice K., 2011, MNRAS, 417, 1928

Forgan D., Rice K., 2012, MNRAS, 420, 299

Forgan D., Rice K., 2013a, MNRAS, 430, 2082

Forgan D., Rice K., 2013b, MNRAS, 432, 3168

Forgan D., Rice K., 2013c, MNRAS, 433, 1796

Forgan D., Rice K., Stamatellos D., Whitworth A., 2009, MNRAS, 394, 882

Forgan D., Rice K., Cossins P., Lodato G., 2011, MNRAS, 410, 994

Forgan D., Parker R. J., Rice K., 2015, MNRAS, 447, 836

Forgan D., Price D. J., Bonnell I., 2017, MNRAS, 466, 3406

Fukagawa M., Tamura M., Itoh Y., Kudo T., Imaeda Y., Oasa Y., Hayashi S. S., Hayashi M., 2006, ApJL, 636, L153

Fukagawa M., et al., 2013, PASJ, 65, L14

Gammie C. F., 1996, ApJ, 457, 355

Gammie C. F., 2001, ApJ, 553, 174

Garcia Lopez R., Natta A., Testi L., Habart E., 2006, A&A, 459, 837

Garufi A., et al., 2013, A&A, 560, A105

Gingold R. A., Monaghan J. J., 1977, MNRAS, 181, 375

Glassgold A. E., Najita J., Igea J., 1997, ApJ, 480, 344

Goldreich P., Lynden-Bell D., 1965, MNRAS, 130, 97

Grady C. A., et al., 2009, ApJ, 699, 1822

Grady C. A., et al., 2013, ApJ, 762, 48

Greenstein J. L., Schmidt M., 1964, ApJ, 140, 1

Gutermuth R. A., et al., 2008, ApJ, 674, 336

Haisch Jr. K. E., Lada E. A., Lada C. J., 2001a, ApJL, 553, L153

Haisch Jr. K. E., Lada E. A., Lada C. J., 2001b, ApJ, 553, L153

Hall C., Forgan D., Rice K., Harries T. J., Klaassen P. D., Biller B., 2016, MNRAS, 458, 306

Harries T. J., Monnier J. D., Symington N. H., Kurosawa R., 2004, MNRAS, 350, 565

Haworth T. J., Harries T. J., Acreman D. M., Bisbas T. G., 2015, MNRAS, 453, 2277

Haworth T. J., et al., 2016, PASA, 33, e053

Hernández J., et al., 2007, ApJ, 662, 1067

Hernquist L., 1993, ApJ, 404, 717

Hubber D. A., Batty C. P., McLeod A., Whitworth A. P., 2011, A&A, 529, A27

Hubeny I., 1990, ApJ, 351, 632

Hubickyj O., Bodenheimer P., Lissauer J. J., 2005, Icarus, 179, 415

Hughes A. M., Wilner D. J., Andrews S. M., Qi C., Hogerheijde M. R., 2011, ApJ, 727, 85

Inutsuka S., 1994, Mem. Soc. Astron. Italiana, 65, 1027

Inutsuka S.-I., 2002, Journal of Computational Physics, 179, 238

Isella A., Tatulli E., Natta A., Testi L., 2008, A&A, 483, L13

Isella A., Natta A., Wilner D., Carpenter J. M., Testi L., 2010, ApJ, 725, 1735

Kley W., Nelson R. P., 2012, ARA&A, 50, 211

Kratter K. M., Murray-Clay R. A., 2011, ApJ, 740, 1

Kratter K. M., Matzner C. D., Krumholz M. R., 2008, ApJ, 681, 375

Kratter K. M., Matzner C. D., Krumholz M. R., Klein R. I., 2010a, ApJ, 708, 1585

Kratter K. M., Murray-Clay R. A., Youdin A. N., 2010b, ApJ, 710, 1375

Kuiper G. P., 1951, Proceedings of the National Academy of Science, 37, 1

Kurosawa R., Harries T. J., Bate M. R., Symington N. H., 2004, MNRAS, 351, 1134

Lada C. J., et al., 2006, AJ, 131, 1574

Lambrechts M., Johansen A., 2012, A&A, 544, A32

Laughlin G., Bodenheimer P., 1994, ApJ, 436, 335

Levermore C. D., Pomraning G. C., 1981, ApJ, 248, 321

Levison H. F., Kretke K. A., Duncan M. J., 2015, Nature, 524, 322

Li Z.-Y., Krasnopolsky R., Shang H., 2011, ApJ, 738, 180

Lin M.-K., 2014, ApJ, 790, 13

Lin M.-K., Papaloizou J. C. B., 2012, MNRAS, 421, 780

Lin D. N. C., Pringle J. E., 1987, MNRAS, 225, 607

Lin C. C., Shu F. H., 1964, ApJ, 140, 646

Lissauer J. J., 1993, ARA&A, 31, 129

Lissauer J. J., Hubickyj O., D'Angelo G., Bodenheimer P., 2009, Icarus, 199, 338

Lodato G., Clarke C. J., 2011, MNRAS, 413, 2735

Lodato G., Rice W. K. M., 2004, MNRAS, 351, 630

Lodato G., Rice W. K. M., 2005, MNRAS, 358, 1489

Lucy L. B., 1977, AJ, 82, 1013

Lucy L. B., 1999, A&A, 344, 282

Lynden-Bell D., Pringle J. E., 1974, MNRAS, 168, 603

Marino S., Casassus S., Perez S., Lyra W., Roman P. E., Avenhaus H., Wright C. M., Maddison S. T., 2015, preprint, (arXiv:1505.06732)

Marois C., Macintosh B., Barman T., Zuckerman B., Song I., Patience J., Lafrenière D., Doyon R., 2008, Science, 322, 1348

Marois C., Zuckerman B., Konopacky Q. M., Macintosh B., Barman T., 2010, Nature, 468, 1080

Masunaga H., Inutsuka S.-i., 1999, ApJ, 510, 822

Masunaga H., Miyama S. M., Inutsuka S.-i., 1998, ApJ, 495, 346

Mathis J. S., Rumpl W., Nordsieck K. H., 1977, ApJ, 217, 425

Mayer L., Lufkin G., Quinn T., Wadsley J., 2007a, ApJ, 661, L77

Mayer L., Lufkin G., Quinn T., Wadsley J., 2007b, ApJ, 661, L77

Mayor M., Queloz D., 1995, Nature, 378, 355

McKee C. F., Ostriker E. C., 2007, ARA&A, 45, 565

McMullin J. P., Waters B., Schiebel D., Young W., Golap K., 2007, in Shaw R. A., Hill F., Bell D. J., eds, Astronomical Society of the Pacific Conference Series Vol. 376, Astronomical Data Analysis Software and Systems XVI. p. 127

Meeus G., et al., 2012, A&A, 544, A78

Meru F., Bate M. R., 2011, MNRAS, 411, L1

Meru F., Bate M. R., 2012, MNRAS, 427, 2022

Monaghan J., 1989, Journal of Computational Physics, 82, 1

Monaghan J. J., 1992, ARA&A, 30, 543

Monaghan J. J., 1997, Journal of Computational Physics, 136, 298

Monaghan J. J., 2002, MNRAS, 335, 843

Monaghan J. J., 2005, Reports on Progress in Physics, 68, 1703

Müller A., van den Ancker M. E., Launhardt R., Pott J. U., Fedele D., Henning T., 2011, A&A, 530, A85

Murante G., Borgani S., Brunino R., Cha S.-H., 2011, MNRAS, 417, 136

Muto T., et al., 2012, ApJL, 748, L22

Naoz S., Farr W. M., Lithwick Y., Rasio F. A., Teyssandier J., 2013, MNRAS, 431, 2155

Nayakshin S., 2010a, MNRAS, 408, L36

Nayakshin S., 2010b, MNRAS, 408, 2381

Nayakshin S., 2011a, MNRAS, 413, 1462

Nayakshin S., 2011b, MNRAS, 416, 2974

Nayakshin S., 2015a, preprint, (arXiv:1502.07585)

Nayakshin S., 2015b, MNRAS, 446, 459

Nayakshin S., 2015c, MNRAS, 448, L25

Nayakshin S., 2015d, MNRAS, 454, 64

Nayakshin S., Fletcher M., 2015, MNRAS, 452, 1654

Nayakshin S., Lodato G., 2012, MNRAS, 426, 70

Nelson R. P., Papaloizou J. C. B., 1994, MNRAS, 270, 1

Nero D., Bjorkman J. E., 2009, ApJ, 702, L163

Oxley S., Woolfson M. M., 2003, MNRAS, 343, 900

Paardekooper S.-J., Mellema G., 2006, A&A, 459, L17

Paczynski B., 1978, Acta Astronomica, 28, 91

Pearce F. R., et al., 1999, ApJ, 521, L99

Pérez L. M., Isella A., Carpenter J. M., Chandler C. J., 2014, ApJL, 783, L13

Pérez L. M., et al., 2016, Science, 353, 1519

Pinte C., Harries T. J., Min M., Watson A. M., Dullemond C. P., Woitke P., Ménard F., Durán-Rojas M. C., 2009, A&A, 498, 967

Pollack J. B., Hubickyj O., Bodenheimer P., Lissauer J. J., Podolak M., Greenzweig Y., 1996, Icarus, 124, 62

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing. Cambridge University Press, New York, NY, USA

Price D., 2005, PhD thesis, PhD Thesis, 2005

Price D. J., 2007, PASA, 24, 159

Price D. J., 2008, Journal of Computational Physics, 227, 10040

Price D. J., 2012, Journal of Computational Physics, 231, 759

Price D. J., Monaghan J. J., 2004, MNRAS, 348, 139

Pringle J. E., 1981, ARA&A, 19, 137

Rafikov R. R., 2005, ApJL, 621, L69

Regly Z., Juhsz A., Sndor Z., Dullemond C. P., 2012, Monthly Notices of the Royal Astronomical Society, 419, 1701

Rice W. K. M., Armitage P. J., 2009, MNRAS, 396, 2228

Rice W. K. M., Armitage P. J., Bate M. R., Bonnell I. A., 2003a, MNRAS, 339, 1025

Rice W. K. M., Armitage P. J., Bate M. R., Bonnell I. A., 2003b, MNRAS, 339, 1025

Rice W. K. M., Armitage P. J., Bonnell I. A., Bate M. R., Jeffers S. V., Vine S. G., 2003c, MNRAS, 346, L36

Rice W. K. M., Lodato G., Pringle J. E., Armitage P. J., Bonnell I. A., 2004, MNRAS, 355, 543

Rice W. K. M., Lodato G., Armitage P. J., 2005a, MNRAS, 364, L56

Rice W. K. M., Lodato G., Armitage P. J., 2005b, MNRAS, 364, L56

Rice W. K. M., Mayo J. H., Armitage P. J., 2010, MNRAS, 402, 1740

Rice W. K. M., Armitage P. J., Mamatsashvili G. R., Lodato G., Clarke C. J., 2011, MNRAS, 418, 1356

Rice W. K. M., Forgan D. H., Armitage P. J., 2012a, MNRAS, 420, 1640

Rice W. K. M., Forgan D. H., Armitage P. J., 2012b, MNRAS, 420, 1640

Rice W. K. M., Paardekooper S.-J., Forgan D. H., Armitage P. J., 2014a, MNRAS, 438, 1593

Rice W. K. M., Paardekooper S.-J., Forgan D. H., Armitage P. J., 2014b, MNRAS, 438, 1593

Rice K., Lopez E., Forgan D., Biller B., 2015, MNRAS, 454, 1940

Ritchie B. W., Thomas P. A., 2001, MNRAS, 323, 743

Rosswog S., 2009, New Astronomy Reviews, 53, 78

Ruge J. P., Wolf S., Uribe A. L., Klahr H. H., 2013, A&A, 549, A97

Safronov V. S., 1960, Annales d'Astrophysique, 23, 979

Sano T., Miyama S. M., 1999, ApJ, 515, 776

Santos N. C., Israelian G., Mayor M., 2004, A&A, 415, 1153

Shakura N. I., Sunyaev R. A., 1973, A&A, 24, 337

Sicilia-Aguilar A., et al., 2006, Astrophysical Journal, 638, 897

Smith R. J., Clark P. C., Bonnell I. A., 2008, MNRAS, 391, 1091

Sod G. A., 1978, Journal of Computational Physics, 27, 1

Spitzer Jr. L., 1942, ApJ, 95, 329

Springel V., Hernquist L., 2002, MNRAS, 333, 649

Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, 726

Srisawat C., et al., 2013, MNRAS, 436, 150

Stamatellos D., 2015, ApJ, 810, L11

Stamatellos D., Whitworth A. P., 2005a, A&A, 439, 153

Stamatellos D., Whitworth A. P., 2005b, A&A, 439, 153

Stamatellos D., Whitworth A. P., 2009, MNRAS, 400, 1563

Stamatellos D., Whitworth A. P., Bisbas T., Goodwin S., 2007, A&A, 475, 37

Terebey S., Shu F. H., Cassen P., 1984, ApJ, 286, 529

Terquem C. E. J. M. L. J., 2008, ApJ, 689, 532

Thi W. F., et al., 2001, ApJ, 561, 1074

Tobin J. J., et al., 2016, Nature, 538, 483

Toomre A., 1964, ApJ, 139, 1217

Verhoeff A. P., et al., 2011, A&A, 528, A91

Von Neumann J., Richtmyer R. D., 1950, Journal of Applied Physics, 21, 232

Vorobyov E. I., 2013, A&A, 552, A129

Vorobyov E. I., Basu S., 2005, ApJL, 633, L137

Vorobyov E. I., Basu S., 2010, ApJL, 714, L133

Vorobyov E. I., Zakhozhay O. V., Dunham M. M., 2013, MNRAS, 433, 3256

Wadsley J. W., Veeravalli G., Couchman H. M. P., 2008, MNRAS, 387, 427

Weidenschilling S. J., 1977, MNRAS, 180, 57

Weidenschilling S. J., 1980, Icarus, 44, 172

Whipple F. L., 1972, in Elvius A., ed., From Plasma to Planet. p. 211

Whitehouse S. C., Bate M. R., 2004, MNRAS, 353, 1078

Whitehouse S. C., Bate M. R., Monaghan J. J., 2005, MNRAS, 364, 1367

Whitworth A., Stamatellos D., Walch S., Kaplan M., Goodwin S., Hubber D., Parker R., 2010, in de Grijs R., Lépine J. R. D., eds, IAU Symposium Vol. 266, IAU Symposium. pp 264–271, doi:10.1017/S174392130999113X

Wick G. C., 1943, Zeitschrift für Physik, 121, 702

Williams J. P., de Geus E. J., Blitz L., 1994, ApJ, 428, 693

Windmark F., Birnstiel T., Güttler C., Blum J., Dullemond C. P., Henning T., 2012, A&A, 540, A73

Wolszczan A., Frail D. A., 1992, Nature, 355, 145

Wright J. T., Marcy G. W., Howard A. W., Johnson J. A., Morton T. D., Fischer D. A., 2012, ApJ, 753, 160

Youdin A. N., 2011, ApJ, 731, 99

Youdin A. N., Goodman J., 2005, ApJ, 620, 459

Zhu Q., Hernquist L., Li Y., 2015, ApJ, 800, 6

van Leeuwen F., 2007, A&A, 474, 653